Mathematical Statistics
Stockholm University

Exponential family models and statistical genetics

Juni Palmgren

Research Report 2000:1

ISSN 0282-9150

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.matematik.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2000:1**,
http://www.matematik.su.se/matstat

Exponential family models and statistical genetics

Juni Palmgren*

January 2000

Abstract

We describe the evolution of applied exponential family models, starting from 1972, the year of publication of the seminal papers on generalized linear models and on Cox regression, and leading up to multivariate (i) marginal models and inference based on estimating equations and (ii) random effects models and Bayesian simulation based posterior inference. By referring to recent work in genetic epidemiology, on semiparametric methods for linkage analysis and on transmission/disequilibrim tests for haplotype transmission we illustrate the potential for the recent advances in applied probability and statistics to contribute to new and unified tools for statistical genetics. We finally emphasise that there is need for well defined post graduate education paths in medical statistics in the year 2000 and thereafter.

KEY WORDS: generalized linear model, partial likelihood, generalized estimating equations, random effects, Bayesian simulation, complex diseases, genetic epidemiology, genetic linkage, genetic association.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: juni@matematik.su.se.

1 Introduction

This article, written more as a narrative than as a comprehensive review, reflects some of my views and experiences as a medical statistician at the door of the new millennium. Although I acknowledge the set of references to be incomplete, it should be comprehensive enough to allow the interested reader to trace the facts and to challenge the arguments.

The advances in molecular biology coupled with modern computer sophistication have fundamentally changed the way in which basic biomedical sciences as well as clinical and epidemiological research approach their art. Old and new data need to be merged, organised and better understood. If one defines medical statistics as 'the scientific contribution to the development and application of tools for the design, analysis and interpretation of empirical studies', then it is clear that our profession should be heavily involved when the new biology and biomedical research practices unfold. Most areas of biomedical research undergo a challenging adaptation process which involves not only the research per se, but also the organisation of research groups and the restructuring of undergraduate and postgraduate education programs. In order for medical statistics to be a serious player on this scene our profession should critically assess its scientific role and clarify the routes to achieve its missions.

The aim of this article is to illustrate the type of challenges that medical statistics are facing. Section 2 describes recent evolution of exponential family models, a branch of statistical science which is central to the field of biomedicine. Section 3 exemplifies the use of exponential family models in modern statistical genetics, the examples involving genetic epidemiology, semiparametric methods for linkage analysis and transmission/disequilibrim tests for haplotype transmission. In section 4 I will briefly return to the issues of education and identity for medical statistics.

2 Exponential family models

Already twenty years ago I found the generalized linear model [1] a useful conceptual framework for communicating the spirit of statistical modelling to biomedical researchers. It provides a unified maximum likelihood framework for regression models with responses from the exponential family. Linear regression for measurements with normally distributed errors, logistic regression for binary responses and Poisson regression for counts are all special cases of the model. The quantitatively inclined applied researcher seems to appreciate the generality of the modelling philosophy and the flexibility of the model specification. The theoretically inclined statistician in turn is often captured by the richness of applied problems that can be handle within

this conceptually simple and unified framework.

Cox's semi-parametric regression models for censored failure time data [2], incidentally published in the same year as the Nelder and Wedderburn paper, has had an equally profound influence on the statistical methodology used in medical research. Cox extends the comparison of life-tables to a general regression setting. Focus is on relative risk parameters and the baseline hazard is treated non-parametrically.

Since the early 70's there has been an escalating effort to deepen the understanding of the properties of these new and practically useful families of non-normal non-linear models. Design issues such as group randomization, litter based toxicology studies, longitudinal studies and family studies have pushed for the need to extend the generalized linear model and the hazard regression model to allow for dependencies within clusters of observations. It was clear at the outset that much of the theory appropriate for multivariate normal linear models could not be extended easily. Instead the normal linear model was to be seen as a special case of an emerging more general statistical modelling framework.

When the models increase in complexity the traditional analytic and numerical methods for model fitting and inference are often inadequate and computationally intractable. The Bayesian model formulation coupled with simulation based inference techniques offer new approaches for model building and evaluation. While computational methods in the form of algorithms and computer programs have always driven the way in which statistical theory has been used in practice, the role of the new simulation based inference tools is fundamentally different. Traditionally the computational tools have evaluated well defined analytic problems, possibly in an iterative fashion. In contrast, the simulation based procedures contribute to the theoretical basis for the model building process and they have added a genuine third leg to the traditional interplay between mathematical theory and real world applications [3]. They do, however, require a new level of computational sophistication on the part of the applied statistician.

The rest of this section elaborates on some of the milestones described above, with the aim to provide a statistical science angle to the methods used in section 3.

2.1 The generalized linear model

The generalized linear model (GLM) is specified through (i) the random variation, i.e. the probability distribution for the observations, and (ii) the systematic variation, i.e. the link function relating the means of the observations to expressions involving the regression paramaters and functions of the covariates. Conditionally on the mean structure the observations are assumed statistically independent. The standard linear regression model is a

GLM with normal distribution and identity link. The log linear model for count data is a GLM with Poisson distribution and log link. The logistic regression model is a GLM with binomial distribution and logit link. These three special cases are useful standard GLM's with attractive theoretical properties [4].

A number of interesting features are connected to maximum likelihood estimation of the regression parameters in a GLM model. Although a full distribution is specified for the observations only the relation between the mean and the variance enters into the score equations for the regression parameters. The score equations are solved by repeatedly solving a set of linear weighted least squares equations. For the normal linear model the algorithm converges in one step. The linear steps of this iteratively reweighted least squares (IRLS) algorithm further suggest that methods appropriate for the normal linear models may be used as approximate tools to evaluate the model fit of the GLM models.

2.2 Cox's partial likelihood regression

Censored failure time data arise naturally in many areas of biomedical research. Early methodology was confined to descriptive life table techniques and to the mathematical formulation of the survival experience over time. Cox's partial likelihood changes the focus to parametric modelling of the survival experience as a function of covariate values, while the baseline time dynamics are treated as a secondary feature and modelled non-parametrically [2],[5]. Central to the model formulation is an assumption of ignorable right censoring, implying that conditionally on covariate values the censored individual is assumed to follow the mean survival experience of those still alive and at risk at the time of censoring.

The partial likelihood ignores the exact time points when failures or censorings occur. Instead, the partial likelihood contribution at each time point when a failure occurs constitutes the conditional probability of observing the failing individual's covariate vector, given all other covariate vectors in the risk set of individuals that could potentially have failed. The properties of the partial likelihood were later formally justified by counting process and martingale theory [6], and light was shed on the formal connection between the partial likelihood, the Mantel-Haenszel test and estimation procedures used in matched case-control analyses and the conditional logistic likelihood based on the hypergeometric distribution [7]. Partial likelihood regression may be viewed as a stratified analysis in which time is controlled for by matching on the risk set at each time point when a failure occurs.

A related argument leads to the relation between the Cox partial likelihood regression and Poisson regression. If a parametric exponential failure time distribution is assumed, then the failure time likelihood is identical to the likelihood formed by treating the failure indicator in a small time interval as

a Poisson variate with the inverse of the hazard as its mean. Allowing the hazard to change between failure times in a piecewise constant fashion results in the Cox partial likelihood and the Poisson likelihood being identical and producing similar inferences for the regression parameters [8].

2.3 Multivariate responses

Logistic regression, Poisson regression and the Cox partial likelihood regression can all be viewed as probability models for a series of binary events [9]. They all share the property that conditional on measured exposures and covariates the responses are assumed independent. Incomplete covariate information, however, induces a need to account for overdispersion as well as residual dependencies. When the sources of heterogeneity are unknown then the additional component of variation may be captured in one single overdispersion parameter [10], [11], [12]. When the data involve identified clusters, e.g. repeated measurements on the same individual or clusters of individuals in families, then a structured model for the within-cluster dependence is called for.

Multivariate normal linear models have been part of the applied statistician's toolbox since the 1930's, but although multivariate binary data, multivariate count data and multivariate censored failure time data arise in many important practical situations, it was not until the late 1980's that Liang and Zeger proposed a general procedure for multivariate generalized linear [13]. An important reason for this slow development is the analytic intractability of non-normal multivariate probability distributions. Further complications arise when covariate effects are linked to the marginal means in a nonlinear fashion, as for the generalized linear model. The beauty of the multivariate normal linear model stems from the fact that the marginal and conditional distributions of a multivariate normal distribution also have normal shape, and when the covariate effects are linked to the marginal means linearly, then the estimation of the regression parameters is independent of the covariance structure. This simplicity is destroyed for multivariate generalized linear models.

During the 1990's joint modelling of the mean and dependence structures for responses measured on a wide variety of scales have been the focus of intense methodological research. The theoretical framework which is emerging essentially follows two main routes: (i) the marginal models and generalized estimating equations and (ii) the random effects models and posterior inference.

2.4 Marginal models and generalized estimating equations

The marginal models focus on the mean structure, and more specifically on the regression parameters linked to the means. The within-cluster dependence is treated as a nuisance, which needs to be accounted for since it affects the power of tests and the precision of the regression estimates. Liang and Zeger [13] constructed generalized estimating equations for the regression parameters by modifying the score equations of a standard generalized linear model. For a generalized linear model based on independent observations the weight matrix in the score equations is diagonal. Liang and Zeger showed that if the mean structure is correctly specified and if the off-diagonal elements in the weight matrix are estimated from the data, then the regression estimates from solving these modified estimating equations are still consistent and normally distributed, with a precision matrix that can be consistently estimated by a robust so called 'sandwich estimator'.

Use of the Liang and Zeger generalized estimating equations for handling residual within cluster dependence does not assume that the 'true' dependence structure is known. One may use a parametrized working weight matrix, the parameters of which are estimated from the residuals at each step in the iteratively reweighted least squares procedure. The closer the specification in the weight matrix is to the 'true' form for the dependence, the more efficient the procedure.

Zhao and Prentice [14] extend the Liang and Zeger procedure for estimating the dependence parameters. They set up two sets of generalized estimating equations jointly, one for the mean parameters and one for the dependence parameters. The first set is based on the observations and involves only the mean parameters, whereas the second set is based on the first order cross products of the observations and involves both the mean and dependence parameters. Since the mean parameters enter both sets of equations these are not independent, implying that the Zhao and Prentice procedure results in different estimates for the regression parameters than the Liang and Zeger procedure. The Liang and Zeger procedure is denoted GEE1, while the Zhao and Prentice extension is denoted GEE2.

Zhao and Prentice further show that the two joint sets of estimating equations constitute the score equations from a quadratic exponential likelihood specified in terms of first and second order dependencies. They term this a pseudo-likelihood since the higher order dependencies are ignored. It is instructive to note that just as the score equations for a generalized linear model based on independent observations only use the relation between the mean and the variance in computing maximum likelhood estimates, the score equations from the Zhao and Prentice multivariate quadratic pseudo-likelihood only uses the first four moments to compute the pseudo-likelihood estimates.

Fitzmaurice and Laird [15] extend the Prentice and Zhao pseudo-likelihood procedure to a full likelihood approach for dependent binary data. They parametrize the joint multinomial likelihood for the multivariate binary observations in terms of the marginal means and the conditional cross-product ratios. This corresponds to a so called 'mixed' parametrisation for an exponential family distribution. In order to solve the resulting set of score equations an inner loop is needed in the iteratively reweighted least squares algorithm in order to express the multinomial cell probabilities in terms of the mean and the dependence parameters. Although there is an one-to-one relationship between these two sets of parameters, no analytic transformation exists. Fitzmaurice and Laird use the iterative proportional scaling (IPS) algorithm for this inner loop. An appealing property of the mixed parametrisation for the exponential family is that the mean and the dispersion parameters are orthogonal to each other, in the sense that the joint covariance matrix for the two sets of parameters is block-diagonal. Parameter orthogonality here implies that inferences for the regression parameters are robust to the properties of the estimated dispersion structure [16].

The various generalized estimating equations desribed above, spanning from the GEE1 via the GEE2 to the score equations for a full multivariate likelihood raise important questions concerning the trade off between bias and precision. If the parametrisation of the full multivariate likelihood holds, then likelihood inference offers asymptotically unbiased and fully efficient estimation for the regression parameters. If, however, the distributional assumptions do not hold, then misspecification could result in biased regression parameter estimates. Note that if the mixed parametrisation does not hold, then the robustness properties induced by parameter orthogonality would be lost. In contrast, if little is known about the multivariate distribution, then the Liang and Zeger GEE1 approach at least provides consistent, albeit less efficient, estimates for the regression parameters. The GEE1 procedure is, however, of no practical use for estimating parameters in the dependence structure. If both the mean and the first order dependencies are of interest, then the GEE2 are preferrable to the GEE1, although the robustness of the mean parameters to misspecification of the dependence structure is partially lost.

Wei, Lin and Weissfeld [17] consider a semiparametric regression model for multivariate right censored failure time data by modelling marginal distributions. They discuss the situation in which two or more distinct failure times are recorded on each individual and the situation in which repetitions of the same kind of event are observed. Each marginal failure time is modelled by a semiparametric Cox model and no specific structure is imposed on the dependence. The regression parameters are estimated by maximizing the failure specific partial likelihoods, resulting in asymptotically consistent normally distributed estimators with a covariance matrix of the 'sandwich type', which may be consistently estimated from the data. The Wei, Lin and Weissfeld marginal model is a multivariate failure time analogue to the GEE1 model for multivariate binary and multivariate count data.

2.5 Random effects models

While the marginal models focus on inference for the fixed regression parameters, the random effects models aim at describing simultaneously both the fixed and the random components of variation. The generalized linear mixed models [18], [19] extend the generalized linear models by adding a set of random terms to the linear predictor. Since no mathematically convenient conjugate distribution is availabe for terms in the linear predictor, the multivariate normal distribution for the random effects is often used as a pragmatic choice. The resulting likelihood is obtained by integrating out the unobserved and often high dimensional vector of random effects. Since the integration cannot be performed analytically a number of approximations have been suggested. Note that the marginal likelihood for the observations is a function of the fixed regression parameters as well as the variance components for the random effects. The individual random effects may be viewed as a set of latent observations and thus the model may be characterized as an incomplete data model. Besides making inferences about the regression parameters and the variance components the purpose of the modelling is often also to make predictions for the individual random effects.

Stiratelli, Laird and Ware [20] elegantly extend to the multivariate binary setting the Laird and Ware [21] random effects model for normally distributed repeated measures. Maximum likelihood estimation is used for the fixed effects and empirical Bayes estimation for the random effects. Since exact solutions are intractable they use an approximation based on the mode of the posterior and they implement the procedure via the EM algorithm. The rationale behind the EM algorithm for likelihood inference for incomplete data was laid out in the seminal paper by Dempster et al [22] and it is briefly described in the next subsection of this paper. By reverting to the mode rather than the mean of the posterior, the E-step and the M-step of the EM algorithm may be merged. Note further that the estimation procedure is doubly iterative and inferences for the fixed and random effects are obtained from the inner loop conditionally on given values for the variance components. The variance components in turn are updated in the outer loop for given fixed and random effects.

Breslow and Clayton [18] derive a penalized partial likelihood solution using Laplace's method for integral approximation to the likelihood. In the multivariate binary setting this procedure, although differently motivated, arrives at the same estimating equations for the fixed and the random effects as those of Stiratelli, Laird and Ware. A third route to similar equations is via an extension of the Henderson best linear unbiased prediction (BLUP) model originally derived for the normal linear random effects model [19],[23].

For right censored failure time data random effects models are referred to as frailty models. Since the Cox partial likelihood regression treats the baseline hazard non-parametrically, there is no intercept in the linear predictor and the notion of overdispersion or heterogeneity in individual risk is in-

trinsically aliased with the baseline hazard itself. Nevertheless, there is a rather extensive literature on so called shared frailty models [24], [25], [26] in which a gamma distributed frailty term acts multiplicatively on the hazard. For clustered failure time data the multiplicative gamma-distributed frailty model has been used, where frailties are divided into additive independent gamma-distributed components [27], [28]. This model, however, results in a complicated likelihood. An altenative is to follow the penalized likelihood approach of Breslow and Clayton [18] and use a multivariate normal distribution for the frailties [29], [30], [31].

The random effects models described above rely on (i) an approximate large sample likelihood solution (ii) computation by an algorithm which has reverted awkward integration to high dimensional matrix inversion (iii) failure to account for undertainty in the estimated variance components when assessing the precision for the estimated fixed and random effects. In a normal linear mixed model the estimates of the regression parameters and the variance components are asymptotically orthogonal. For exponential family mixed models this orthogonality property does not hold, and the uncertainty in the one set of parameters should be incorporated in the estimated precision for the other set. The Bayesian model formulation and simulation based posterior inference constitute an alternative to the large sample iterative procedures.

2.6 Bayesian simulation based inference

Random effects models are appropriate for many real world problems, but they are difficult to fit using traditional statistical tools. New simulation based techniques to draw inferences from these models can be viewed as extensions to the EM algorithm for maximum likelihood estimation of incompletely observed data, or they can be derived from a Bayesian perspective. Here the incompleteness refers to the unobserved random effects. Other additional incomplete features of the data may naturally be incorporated into the model.

Heuristically the EM algorithm works by first filling in the missing data, then estimating the parameters from the completed data and then re-estimating the missing data using the updated parameter values. Formally the EM algorithm maximizes the observed data likelihood by iteratively maximizing the complete data likelihood [22]. Each iteration consists of two steps. The E step computes the expectation of the complete data log likelihood over the predictive distribution for the missing data, given the observed data and the current parameter estimates. The M step maximizes the ensuing conditional expected complete data log likelihood using the same maximum likelihood routine as would be used for a complete data likelihood. The EM algorithm is easy to monitor but the convergence may be slow, the rate of convergence being proportional to the fraction of observed data relative to the complete data [32]. The E step is often itself intractable, as is the case for the random

effects models with a non-conjugate distribution, for which the E step involves the same intractable integral as the observed data likelihood.

One method to avoid the intractable E step is to evaluate the mode rather than the mean. This was discussed in the previous subsection. Another approach involves drawing samples from the predictive distribution of the missing data given the observed data and the current parameter values and to calculate the Monte Carlo mean. This Monte Carlo EM method maximizes the observed data likelihood, and is thus a large sample iterative technique.

A conceptually different solution involves replacing both the E step and the M step by successive draws from respectively the predictive distribution for the missing data given the observed data and current parameter value, and from the distribution for the model parameters given the completed data. Since this data augmentation approach [33] assumes that all model parameters, including the regression parameters and the variance components are random, it has a Bayesian flavour and is conceptually different from a large sample iterative EM solution for maximizing the observed likelihood. Given certain regularity conditions the successive draws in this data augmentation procedure will eventually converge to the joint posterior distribution for the missing values and the parameters. Simulated marginal posterior distributions are thus available for the regression parameters, the random effects parameters and the variance components, respectively. If priors with very large variances are used for the regression parameters and the variance components, then the posterior means for the regression parameters closely correspond the the maximum likelihood estimates. Note, however, that that posterior credible intervals for the regression parameters and the random effects incorporate the uncertainty in the estimated variance components. This resolves the problem inherent in the large sample iterative approach, in which the inferences for the regression parameters and the random effects are made conditional on the variance components being fixed.

Iterative simulation techniques date back at least to Metropolis et al [34] in the physical sciences. Tanner and Wong [33] were influential in introducing these methods and related theory and examples to statistical science. The Gibbs sampler is a useful special case of the Metropolis algorithm in which the missing values and the model parameters are partitioned into sets, and in one iteration of the Gibbs sampler all full conditional distributions are sampled in turn. The Gibbs sampler is useful when the conditional distributions are easy to sample from.

The several recent books on Bayesian computation indicates that this new methodology for statistical inference is here to stay [35], [36], [37], [38]. The simulation based algorithms provide an attractive, coherent and flexible inference framework for a large set of models that could not be handled by traditional tools. The models naturally incorporate incomplete data structures as well as prior information from external sources. The fitting procedure is, however, computationally intensive, the convergence properties of the different samplers are difficult to assess, and the sensitivity to the various

model assumption are not transparent. More experience of these methods is needed.

Rubin [32] gives a lucid overview of computational aspects of analysing random effects models and Zeger and Karim [39] provide a full Bayesian posterior analysis of a Poisson log linear random effects model.

3 Statistical genetics

Due to the enormous amount of information provided by today's genome projects the field of genetics is experiencing an outburst of empirical exploration of its theoretical roots: the actual identification of the chromosomal loci underlying phenotypic variation. The practical implications of this development for medical practices is likely to be enormous.

There is a growing literature with statistical orientation on the topics of segregation ratios, population frequencies, genetic linkage, allelic association and continuous and quasi-continuous traits scattered in numerous books and periodicals devoted to the various overlapping branches of genetics such as medical genetics, population genetics, quantitative genetics, behavioural genetics, molecular genetics and genetic epidemiology. The flow of recent books on statistical genetics reflect attempts at structuring and unifying the methodological issues in the field [40], [41], [42], [43], [44], [45], [46].

Below we first give a brief account of the early controversy between Mendelian and quantitative genetics, resolved by a unified view on the laws of segregation for traits measured on different scales. We emphasise the methodological implications from reverting interest from single gene disorders to complex diseases, and finally claim through examples involving genetic epidemiology, multipoint linkage and linkage disequilibrium methods that the emerging multivariate statistical modelling framework described in section 2 has the potential to contribute new and unified tools for genetic outcomes measured on a wide variety of scales.

3.1 Theoretical roots

Mendelian genetics in 1900 was concerned with inheritance of discrete characters such as purple vs white flower color, blood-group, eye-color, wrinkled vs smooth seeds. The mechanism of inheritance can be observed through the 'Mendelian ratios' only when a gene difference at a single locus gives rise to a readily detectable discrete trait difference.

The Mendelian theory appeared to be in sharp contrast to an independent

branch of empirical genetics begun earlier by Francis Galton [47] who concentrated on continuously varying characters. A series of debates ensued between the Mendelians led by William Bateson and the Biometricians led by Karl Pearson. The major issues were whether discrete characters have the same hereditary and evolutionary properties as continuously varying characters. The Mendelians view evolution as arising from genetic mutations with large effects while the Biometricians viewed evolution as the result of natural selection acting on continuously distributed traits [45].

Mendel himself had suggested an explanation for how variation in continuous characters could be maintained by the independent segregation of multiple factors. The British mathematician Udny Yule gave formal proof for this idea in 1902 [48]. Unfortunately, at that point in time the only thing that the Biometric and the Mendelian schools could publicly agree on was the incompatibility of Mendelian genetics and the inheritance of continuous characters. It was Ronald Fisher who in 1918 wrote the classic paper entitled 'The correlation between relatives on the supposition of Mendelian inheritance', which reconciled the two schools [49]. The standard model for inheritance of quantitative trait values assumes a large number of loci that act independently and additively, each with a small effect on the trait, and each following the laws of Mendelian transmission.

The importance of statistics in human genetics has a long history. Karl Pearson and Ronald Fisher, two of the pioneers of statistics in the early 20th century, were both involved in genetics at some point in their career. Francis Galton provided the empirical motivation for Karl Pearson's formal development of the theory of regression and correlation [50]. Fisher's 1918 paper introduced the concept of variance-component partitioning and Wright introduced path analysis in 1921 [51].

3.2 Complex diseases

The Human Genome Project and the DNA-sequencing of many other organisms have reveiled a modular structure of the genome, building up from nucleotides to codons to gene families and other higher order structures. It is the duplication of whole genes or clusters of genes with subsequent modification which provides the material for inherited phenotypic variation.

Early work in modern genetics has been dominated - with much success - by the study of single gene disorders. This typically involves identification of large multiply affected pedigrees, estimation of penetrances and modes of transmission by segregation analysis and 'parametric' linkage analysis. A chromosomal region is thus identified in which the causative gene must lie. Subsequent more refined analysis narrows down the region to a size small enough for exhaustive search with molecular biological techniques.

Recent interest has turned to diseases of more complex aetiology, such as

diabetes, multiple sclerosis, rehrumatoid arthritis, cardiovascular disease, asthma, hypertension and psychiatric illnesses, which are assumed to involve more than one causative gene. The complex transmission reflecting the actions and interactions of multiple genetic and environmental factors requires development of new methodology. Studies based on large numbers of simple pedigrees ascertained from population-based samping frames are becoming commonplace, and established methods for linkage analysis are giving way to methods based on affected pairs of siblings and to the study of linkage disequilibrium using population and family-based cases and controls.

It is of interest to note that recent identification of quantitative trait loci (QTL) from DNA-level data suggests that the influence on quantitative trait variation could stem from one or two alleles with strong effect, and many alleles with minor effect at the same locus. The number of QTL's may thus be small, or at least finite. This suggests a unified way in which discrete and quantitative traits are produced at the gene level, and is changing the perspective on the mechanism of quantitative trait inheritance [52].

Trait variation, which may be measured on a continuous, binary or censored age-at-onset scale is thus determined by observed and unobserved variation in multiple genetic and environmental factors. Power to map new disease genes is increased if chromosomal regions and environmental factors already identified as linked to the disease are accounted for. One cannot, however, assume that all genetic and environmental factors affecting disease susceptibility have been identified. Strong residual dependence between phenotypes of family members often remains [53], which on one hand renders invalid the traditional within-family assumption of conditional independence for the trait given the genotype, and on the other hand may contain important information on the magnitude and character of as yet unidentified sources of variation.

The multivariate exponential family models are useful for modelling residual variation in a variety of settings, including within family dependence and dependence in marker expression and in recombination counts along chromosomal segments. We present recent contributions to genetic epidemiology, multipoint linkage analysis and linkage disequilibrium testing in which the methods described in section 2 are utilized.

3.3 Genetic epidemiology

Historically, epidemiology and genetics have been different in their outlook. Epidemiology has focussed on effects of environmental factors including age and gender. The environmental exposures have been observable, although possibly measured with error. The focus of genetic studies on the other hand has been on factors such as the Huntington disease gene or breast cancer genes, and traditionally the putative disease genes have been latent and not directly observable in individuals. These differences have resulted in different

conceptual orientations in defining risk factors, in choosing study designs and in the statistical methods used to extract information. Still, epidemiology and genetics are intrinsically connected in that both share the same mission of trying to understand the aetiology of human diseases, whether genetic or environmental. With molecular data the differences in outlook are beginning to disappear. Contemporary epidemiologic studies often use biological markers, including candidate genes if available, and genetic studies increasingly consider influences of environmental factors in their penetrance functions.

Burton et al [54] use a Bayesian model and Gibbs sampling to fit a generalized linear mixed model to binary phenotypes in nuclear families. They illustrate their model on a study of the genetics of atopic disease. Some two hundred families consisting of two parents and at least two children were sampled from the population of eligible families in the town of Busselton in Western Australia. Individuals were defined as atopic if they satisfied any one of a series of standard criteria. A generalized linear mixed model was specified with atopy as the binary response, a logit link function, fixed effect terms representing each of eight age-gender groups and three random effect terms representing an additive polygenic effect, an effect of common family environment and an effect of common sibling environment. The model was further extended to look for linkage disequilibrium with alleles 3 and 4 of a microsatellite marker located on chromosomal region 11q13. Assumptions of Hardy-Weinberg equilibrium, random mating and random ascertainment underly the model specification. The results described in the paper provide moderately strong evidence for an additive genetic variance component. Although the microsatellite marker in question has shown linkage to quantitative phenotypes related to atopy, no evidence of linkage disequilibrium with the binary phenotype 'atopy' was shown in this study.

3.4 Semiparametric methods for multipoint linkage

Zhao et al [55], [56] propose a semiparametric model for two-point and multipoint linkage analysis and they use the method on breast cancer family data. They emphasise that the approach handles binary, continous, or censored failure time phenotypes in a unified fashion, and that it is applicable to different family structures including extended pedigrees, nuclear families, sib pairs, affected relative pairs or mixtures of these family structures.

The two-point model estimates the recombination fraction between the putative disease locus and one marker at a time. The multipoint model uses multiple marker loci simultaneously and is expected to be more efficient than the two-point analysis. Multipoint linkage is of interest particularly as the Human Genome Map is being constructed, offering many genetic markers for the mapping of complex traits.

For each family member data is available on the phenotype, on covariates including candidate genes and environmental factors and on markers at a

number of loci. Besides the observed data a latent putative disease genotype is specified, with alleles at the putative disease locus of the mutant or wild type.

The parametric component of the proposed semiparametric model assumes known form for the penetrance function and specified allele frequency for the putative disease genotype. In addition Hardy-Weinberg equilibrium and Mendelian transmission is assumed for the putative disease gene. Residual dependence is acknowledged for phenotypes of members within a family, but no distributional form is assumed for the joint phenotype distribution. An estimating equation approach is used to estimate recombination fractions and to make inferences about the position of putaitve disease genes. Dependencies between the counts of recombinants for given counts of informative meiosis on different chromosomal segments are built into the model. The dependencies may reflect interference and uncertain marker order. Paternal and maternal meiosis may be accounted for separately. Here the segments constitute windows which are set to move over the chromosome and an estimate of the average recombination fraction is obtained. The possibility to account for dependencies between marker segments explains the increased power for the multipoint model relative to the two-point model. Zhao et al [56] discuss in detail the assumptions concerning higher order dependencies as well as the pros and cons of using the GEE1 and GEE2 approaches.

3.5 Transmission/disequilibrium tests for haplotype transmission

Traditional linkage studies explore the fact that genetic markers near to the disease susceptibility gene tend to be inherited together with the disease susceptibility gene itself. The observation of recombination events identifies if the disease gene lies on a particular part of a chromosome or not. Usually linkage studies will only be able to locate genes to an accuracy of a few cM, which leaves a very large region of DNA to be sequenced. Even with the increased power from multipoint linkage studies a very large number of multiply affected families would be needed to narrow the region.

As an alternative to traditional linkage studies the mapping of disease susceptible genes to smaller chromosomal regions may be possible by considering 'a very large' family, i.e. the population. Population based studies comparing allele frequencies in cases and controls have, however, been criticized as prone to false positive findings due to population admixture, i.e. unidentified population heterogeneity involving varying allele frequency and varying disease risk in the latent subgroups. To overcome the confounding problem induced by population admixture family based case-control study designs are used. These are based on genotyping of cases and both their parents and use the non-transmitted alleles as family based controls [57], [58].

Clayton and Jones [59], [60] develop a formal statistical framework for transmission/disequilibrium tests (TDT) to detect association between polymorphic markers and categorical or quantitative traits. Emphasis is on marker haplotypes formed by several adjacent loci and the methodology is targeted for fine mapping using a set of diallelic SNP-markers in an identified candidate region.

First a general haplotype relative risk model is defined, in which the relative risk of a heterozygote genotype is defined to be intermediate between the two homozygote genotypes, the exact position being determined by the particular form of a monotone increasing link function. The TDT test can be derived as a score test for the hypothesis that the haplotype relative risk parameters are all unity. This is equivalent to a score test based on the hypergeometric distribution in a matched case-control study. As implied by the relationships between models discussed in section 2.2 the full toolbox for conditional logistic regression is available [7], [61]. Information on individual specific and family specific environmental factors may be incorporated into the analysis. Clayton and Jones emphasize that also for quantitative traits it is useful to condition on the offspring trait value and the parental genotype and to treat transmission as the random response. A conditional likelihood is formed which is parametrized in terms genotype specific deviations in average trait value from the overall trait mean. The likelihood contribution is defined as the probability that conditional on the offspring phenotype value the genotype of the affected offspring is transmitted rather than any other of the possible genotypes.

With increasing number of loci the marker haplotype polymorphism increases rapidly as well as the number of haplotype relative risk parameters under the alternative. If the model under the alternative is left unspecified a global test will lack power. For binary or discrete traits, the log linear model provides a useful conceptual framework for defining disease-marker associations [62]. The practice of using first order marker associations is simple, but it is questionable in that the most informative marker is not necessarily the one that is physically closest to the disease susceptibility locus. It is worth noting that the closer one is to the time of the disease mutation the higher the order of expected haplotype association. If one is close in time to the mutation a test of the null hypothesis against a very-high-order alternative is powerful. With increasing distance in time from the mutation the high-order effect is rapidly diluted by recombination. As an alternative to searching the hierarchy of tests, which would ultimately lead to a multiple testing problem, Clayton and Jones propose a random effects alternative. This is based on assuming the haplotype relative risks parameters to be random and generated by a multivariate normal model with a covariance matrix specified in terms of measurable haplotype similarity and a single hyperparameter determining the extent of the association. Haplotype similarity is defined by the location and length of the longest contiguous chromosomal segment in the candidate region over which there is identity by state.

The general TDT model is further extended to account for incomplete trans-

mission information. This may arise if parents are unavailable for typing, or if the information content in the SNP's does not allow phase determination. Restricting attention to cases where parental genotypes may be inferred from additional offspring genotypes has been shown to be prone to bias [63]. Instead Clayton describes how standard statistical models for incomplete data [22], [64] may be used for model specification and to obtain a modified score and information for the incompletely observed data. The authors point out that a hierarchical Bayes model and simulation based posterior inference are well suited for estimation and model assessment in this general haplotype relative risk model.

4 Medical statistics 2000

While computer science and bioinformatics provide methods for storage, organisation and retrieval of the accumulating molecular information together with algorithms for pattern recognition and prediction, statistical science is concerned with the quantification of sources of variation and uncertainty, and with the assessment of power and robustness for methods used in the design, analysis and interpretation of empirical studies.

The aim of this paper has been to illustrate from one specific angle how the frontier in applied probability and statistics may contribute important new research tools to empirical biomedicine. A somewhat fragmented glimpse is provided of a scenario which is still in its infancy. It should, however, be clear that the development of new statistical tools for empirical research stands on three legs: (i) innovative use of mathematics, probability theory and statistical inference theory, (ii) good understanding of the often complex biological phenomena and (iii) understanding of modern rather sophisticated computational tools.

In order for medical statistics to fulfil its role as a scientific player on the emerging scene for empirical biomedicine, new post graduate education paths need to be developed, based on mathematics and theoretical statistics and firmly integrated with biology, medicine and computer science. A challenge for the year 2000 and the years to follow!

REFERENCES

[1] Nelder JA, Wedderburn RWM. Generalized linear models. Journal of the Royal Statistical Society, Series A 1972;135:370-84.

[2] Cox DR. Regression models and life tables (with discussion). Journal of the Royal Statistical Society, Series B 1972;43:187-220.

[3] Efron B, Tibshirani R. Computer-intensive statistical methods. In Armitage P and David HA. Eds. Advances in Biometry. New York: Wiley. 1996.

[4] McCullagh P, Nelder JA. Generalized linear models. London: Chapman and Hall. 1989.

[5] Cox DR. Partial likelihood. Biometrika 1975;62:269-76.

[6] Andersen PK, Gill RD. Cox's regression models for counting processes: a large sample study. Annals of Statistics 1982;10:1100-20.

[7] Breslow NE, Day NE. Statistical methods in cancer research. Vol I. The analysis of case-control studies (IARC Scientific Publications No. 32). Lyon: International Agency for Research on Cancer. 1980.

[8] Frome EL. The analysis of rates using Poisson regression models. Biometrics 1983;39:665-74.

[9] Clayton D. Some aproaches to the analysis of recurrent event data. Statistical Methods in Medical Research 1994;3:244-262.

[10] Wedderburn RWM. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika 1974;61:439-47.

[11] Williams DA. Extra-binomial variation in logistic linear models. Applied Statistics 1982;31:144-148.

[12] Breslow NE. Extra-Poisson variation in log linear models. Applied Statistics 1984;33:38-44.

[13] Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:11-22.

[14] Zhao LP, Prentice RL. Correlated binary regession using a quadratic exponential model. Biometrika 1989;77:642-48.

[15] Fitzmaurice GM, Laird NM. A likelihood based method for analysing longitudinal binary responses. Biometrika 1993;80:141-51.

[16] Cox DR, Reid N. Parameter orthogonality and approximate conditional

inference. Journal of the Royal Statistical Society, Series B 1987;49:1-39.

[17] Wei LJ, Lin DJ, Weissfeldt L. Regression analysis of multivariate incomplete failure time data by modelling marginal distributions. Journal of the American Statistical Association 1989;84:1065-73.

[18] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 1993;88:9-25.

[19] McGilchrist CA. Estimation in generalized mixed models. Journal of the Royal Statistical Society, Series B 1994;56:61-69.

[20] Stiratelli R, Laird NM, Ware H. Random-effects models for serial observations with binary response. Biometrics 1984;40;961-71.

[21] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963-74.

[22] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete observations via the EM algorithm. Journal of the Royal Statistical Society, Series B 1977, 39:1-38.

[23] Robinson GK. That BLUP is a good thing: the estimation of random effects. Statistical Science 1991;6:15-51.

[24] Clayton D. Cuzick J. Multivariate generalizations of the proportional hazards model (with discussion). Journal of the Royal Statistical Society, Series A 1985:148;82-117.

[25] Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. Biometrics 1992;48:795-806.

[26] Andersen PK, Borgan O, Gill RD, Keiding N. Statistical models based on counting processes. Heidelberg: Springer-Verlag. 1993.

[27] Petersen J, Andersen PK, Gill RD. Variance component models for survival data. Statistica Neerlandica 1996;50:193-211.

[28] Korsgaard IR, Andersen AH. The additive genetic gamma frailty model. Scandinavian Journal of Statistics 1998;25:255-69.

[29] Therneau TM, Grambsch PM. Penalized Cox model and frailty. S-Plus frailty function documentation.

[30] Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. Research Report. University of Copenhagen, Department of Biostatistics 1999.

[31] McGilchrist CA. REML estimation for survival models with frailty. Bio-

metrics 1993;49:221-225.

[32] Rubin DR. Computational aspects of analysing random effects/longitudinal model. Statistics in Medicine 1992;11:1809-1821.

[33] Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association 1987;82:528-50.

[34] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller E. Equations of state calculations by fast computing machines. Journal of Chemical Physics 1953;21:1087-91.

[35] Tanner MA. Tools for statistical inference. Mathods for the exploration of posterior distributions and likelihood functions. (second edition). New York: Springer. 1993.

[36] Gelman A, Rubin DF, Carlin J, Stern H. Bayesian data analysis. London: Chapman and Hall. 1995.

[37] Schafer JL. Analysis of incomplete multivariate data. London: Chapman and Hall. 1996.

[38] Gilks WR, Richardson S, Spiegelhalter DJ. Markov chain Monte Carlo in practice. London: Chapman and Hall. 1996.

[39] Zeger SL, Karim MR. Generalized linear models with random effects; A Gibbs sampling approach. Journal of the American Statistical Association 1991;86:79-86.

[40] Khoury MJ, Beaty TH, Cohen BH. Fundamentals of genetic epidemiology. Oxford: Oxford University Press. 1993

[41] Speed T, Waterman MS. Genetic mapping and DNA sequencing. New York: Springer. 1996.

[42] Falconer DS, Mackay TFC. Introduction to quantitative genetics (fourth edition). Harlow: Longman. 1996.

[43] Lange K. Mathematical and statistical methods for genetic analysis. New York: Springer. 1997.

[44] Sham PC. Statistics in human genetics. London: Arnold. 1998.

[45] Liu BH. Statistical genomics. Linkage, mapping and QTL analysis. New York: CRC Press. 1998.

[46] Lynch M, Walsh B. Genetics and analysis of quantitative traits. Massachusetts: Sinauer Association, Inc. 1998.

[47] Galton F. Natural Inheritance. London: Macmillan. 1889.

[48] Yule GU. Mendel's laws and their probable relations to intra-racial heredity. The New Phytologist 1902;1:193-207.

[49] Fisher RA. The correaltion between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 1918;52;399-433.

[50] Stigler SM. Francis Galton's account of the invention of correlation. Statistical Science 1989;4:73-9.

[51] Wright S. Correlation and causation. Journal of Agricultural Research 1921;20:557-85.

[52] Weiss KM. Introduction to the symposium on Variation in the human genome. Eds Chadwick D, Cardew G. New York: Wiley. 1996.

[53] Hopper J. Variance components for statistical genetics: applications in medical research to characterics related to human disase and health. Statistical Methods in Medical Research 1993;2:199-223.

[54] Burton PR, Tiller KJ, Gurrin LC, Cookson WOCM, Musk AW, Parmer LJ. Genetic variance components analysis for binary phenotypes using generalized linear mixed model (GLMMs) and Gibbs sampling. Genetic Epidemiology 1999;17:118-40.

[55] Zhao LP, Quiaoit F, Hsu L, Aragaki C. An efficient, robust, and unified method for mapping complex traits (I): Two-Point linkaga analysis. American Journal of Medical Genetics 1998;77:366-83.

[56] Zhao LP, Quiaoit F, Aragaki C, Hsu L. An efficient, robust, and unified method for mapping complex traits (II): Multipoint linkaga analysis. American Journal of Medical Genetics 1998;79:48-61.

[57] Falk CT, Rubinstein P. Haplotype relative risk: an easy reliable way to contruct a proper control sample for risk calculation. Annals of Human Genetics 1987;51:227-33.

[58] Terwillinger J, Ott J. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. Human Heredity 1992;42:337-46.

[59] Clayton D. A generalization of the Transmission/Disequilibrium Test for uncertain haplotype transmission. American Journal of Human Genetics 1999;65:1170-77.

[60] Clayton D, Jones H. Transmission/Disequilibrium Tests for extended marker haplotypes. American Journal of Human Genetics 1999;65:1161-69.

[61] Self SG, Lonton G, Kopecky KJ, Liang KY. On estimating HLA/disease association with application to the study of aplastic anemia. Biometrics 1991;47:53-61.

[62] Chiano M, Clayton D. Fine genetic mapping using haplotypes and the missing data problem. Annals of Human Genetics 1998;62:55-60.

[63] Curtis D, Sham PC. Using risk calculation to implement an extended relative pair analysis. Annals of Human Genetics 1994;58:151-62.

[64] Little RJA, Rubin DB. Statistical analysis with missing data. New York: Wiley. 1987.