An Overview of Statistical Models and Statistical Thinking

David Firth

Department of Statistics, University of Warwick
and
ESRC National Centre for Research Methods

ESRC Oxford Spring School, 2007-03-27

Preface

Models are central to (almost) all statistical work.

This short course aims to give an overview of some of the most prominent statistical models, and associated methods for inference, interpretation and criticism, used in social research.

The focus will be on models of dependence.

Much will be treated very superficially. The aim is an overview: more detailed understanding will require further reading.

Computer lab sessions will be used to illustrate some of the less standard models/methods.

Plan

Part I: Models

Part II: Inference

Part III: Diagnostics

Part IV: Interpretation

Part V: Things to Worry About

Part I: Models

Purposes

Some General Principles

Types

Part II: Inference

General Notions

Likelihood

Estimating Equations

Simulation, Bootstrap

Part III: Diagnostics

General Principles

Residuals

Lack-of-fit Tests

Influence

Part IV: Interpretation

Precision and Significance

Parameterization

Conditional Independence, Graphical Models

Causality

Part V: Some Things to Worry About

(briefly...)

Part I

Models

Purposes of Modelling

e.g., multiple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i$$

Explanation: How does (the mean of) y change with x and z?

Prediction: What is the expected value of y, and how much variation is there in the value of y, at a particular combination of x and z?

Accounting: What is the average value of y in the population? [This is a special application of prediction, really.]

Some General Principles

To be useful, a statistical model should:

- embody the research questions of interest, via (functions of) parameters, (conditional) independence relationships, etc.
- take account of research design (e.g., cluster sampling)
- reflect all *important* structure — systematic and haphazard variation — in the data
- not be more complicated than necessary

What is a statistical model?

A statistical model is a family of probability distributions.

For example, $N(\mu, \sigma^2)$ is a distribution. The *parameters* μ and σ together index a *family* of such distributions: each different (μ, σ) combination corresponds to a different normal distribution.

In the linear model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + e_i,$$

if we assume $e_i \sim N(0, \sigma^2)$, the family is a set of normal distributions indexed by $\beta_0, \beta_1, \beta_2, \sigma$.

Univariate/Multivariate

This terminological distinction varies between disciplines.

In discussing models of dependence, **univariate** means that there is a single response or outcome variable (y). The number of explanatory variables (predictor variables, covariates) may be none, one, or many.

By contrast, **multivariate** means modelling the distribution of two or more response variables $(y_1, y_2, \ldots)$ jointly. Again, the number of explanatory variables may be none, one or many.

Time Series and Panel Data

Here the univariate/multivariate distinction is less clear.

Same variable (e.g., income) measured at a series of time-points: y_t is the value at time t.

Time series: typically means a long series of observations, e.g., monthly exchange rate over 20 years. **Univariate** time-series models relate to the study of such a series in isolation; **multivariate** if that series is related to others, such as monthly interest rate for the same time-points, etc.

Panel data: several short time series, one for each unit of analysis. Aim is often to relate mean level or trend to explanatory variables. Sometimes univariate models suffice; sometimes multivariate models are needed.

Level of Assumptions

The family of distributions may be parametric, non-parametric, or ‘semi-parametric’. Broadly speaking:

parametric: the family of distributions is fully specified, up to a small number of unknown parameters. For example, $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$.

semi-parametric: some aspects of the distribution are described by parameters, but others are left unspecified. For example, $E(y) = \beta_0 + \beta_1 x$, $\mathrm{var}(y) = \sigma^2$; here the distributional *shape* is unspecified.

non-parametric: no distributional assumptions other than known consequences of the sampling scheme.

A **nonparametric** ‘model’ is really no more than a description of the sampling scheme and a definition of the quantity of interest. This has appeal in regard to robustness to failure of assumptions (no assumptions are made!).

In practice, nonparametric models are used only with simple research designs involving one or two variables, for which there is well-developed statistical theory.

Parametric models, backed up by thorough diagnostic checking of assumptions, are much more widely used in social research. **Likelihood** provides a unified framework for inference.

Semi-parametric models provide a middle road, in situations where there are parameters of primary interest and other aspects that do not need to be modelled in detail. This admits the use of ‘robust’ methods of inference (partial likelihood, quasi-likelihood, GEE, etc.) which work for the primary parameters under weak conditions on the un-modelled aspects. Examples include:

- the Cox proportional hazards model for duration data
- overdispersed models for binomial/count data
- marginal models for clustered/panel data.

Linearity

The ‘linear’ in ‘linear model’ refers to *linearity in the parameters*.

Thus

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

is a linear model for the mean of y, but

$$E(y) = \beta_0 + \beta_1 x^{\beta_2}$$

is non-linear.

Linear models are easier to fit, have nicer statistical theory, and are simpler to interpret.

Transformation to Linearity

Sometimes a nonlinear model can be *transformed* into a linear model, by applying a suitable mathematical function to y and/or explanatory variables.

A common case is the transformation of a multiplicative model such as

$$y_i = \beta_0 x_i^{\beta_1} z_i^{\beta_2} e_i$$

(with $E(e_i) = 1$) into the linear form with additive error

$$\log y_i = \beta_0^* + \beta_1 \log x_i + \beta_2 \log z_i + e_i^*.$$

Generalized Linear Model

Non-linear models of the particular form

$$g[E(y)] = \beta_0 + \beta_1 x + \beta_2 z + ...$$

still (typically) require iterative fitting, but they inherit much of the nice statistical theory and ease of interpretation.

Here g is a specified, smooth, monotone transformation known as the **link function**. On the scale of the link function, $E(y)$ is described by the familiar kind of linear predictor.

GLM versus Response Transformation

e.g., log-link generalized linear model

$$\log E(y) = \beta_0 + \beta_1 x$$

versus log-transformation of the response variable

$$E(\log y) = \beta_0 + \beta_1 x.$$

The second is a linear model for the mean of $\log y$. This may or may not be appropriate; e.g., if y is income, perhaps we are really interested in the mean income of population subgroups, in which case $E(y)$ and not $E(\log y)$ is the right thing to model.

The second also has technical problems if any $y = 0$.

Additivity

Additivity refers to the contributions of explanatory variables.

Linear models are not necessarily additive, since they can involve interaction terms (the effect of x may be different at different levels of z, and *vice versa*).

A **generalized additive model** has, for example,

$$g[E(y)] = \beta_0 + f_x(x) + f_z(z)$$

in which f_x, f_z are functions, possibly but not necessarily parametric.

Often in practice the f functions are specified as cubic splines, because they offer a good compromise between flexibility and economy of representation.

Additivity in such models can be relaxed in structured ways, e.g.,

$$g[E(y)] = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + f_x(x)$$

has linear interaction between x and z, combined with a flexible non-trend part f_x in the dependence on x.

Care is needed in interpretation. In the example above, for example, the function $f_x(x)$ should be trendless in order to allow the usual interpretations for β_0 and β_1.

The *R* package `mgcv` provides very good facilities for fitting models involving functions like $f_x(x)$, $f_z(z)$, by penalized maximum likelihood methods.

Additive models with flexible functional dependence allow exploration of the shape of ‘main-effect’ aspects of dependence.

At the opposite extreme is flexible exploration of *interaction*.

In some contexts, interaction effects may be far from linear.

One tool for exploring ‘unstructured’ interaction effects is the **regression tree** model (also known as **recursive partitioning**).

Regression trees can be hard to interpret, and can be very non-smooth (hence unrealistic). But they are a useful exploratory device for complex dependences.

Response-variable type

The following response (y-variable) types are commonly encountered:

- ‘continuous’ (measurements etc.)
- counts (of events etc.)
- categorical
  - binary
  - nominal
  - ordinal
- durations (survival or event-history data)

Mixture of continuous and discrete (e.g., where response is either zero or a positive amount of something) sometimes occurs, and demands special care.

Response type: continuous

If a model is required for $E(y)$, consider GLM with suitably-chosen link function.

Alternatively, use a linear model, possibly after a non-linear transformation of y.

GLM has advantage of allowing variance to depend on mean in a specified way. For example, with homogeneous multiplicative errors, variance = $\phi[E(y)]^2$.

In a GLM (or GAM) the link function is chosen to achieve linearity (additivity) of the right hand side.

Often (but not necessarily) this means linking the mean in such a way that $g[E(y)]$ can take any real value. For example, if $E(y) > 0$, $g(\mu) = \log \mu$ will often be a candidate.

Response type: counts

e.g., numbers of arrests made by different police forces in different time periods.

Interest is most often in the **rate** of occurrence per unit of exposure, where ‘exposure’ might be amount of time, population at risk, person-hours of effort, or a composite.

Most natural starting point is a Poisson model with log link: $y_i \sim \text{Poisson}(\mu_i)$, with, say,

$$\log \mu_i = \log t_i + \beta_0 + \beta_1 x_i + \beta_2 z_i$$

where t_i is the known exposure quantity for the ith count. The term $\log t_i$ here, with no unknown coefficient attached to it, is called an *offset*. It ensures that the other effects are all interpretable as rate-multipliers.

Response type: binary

If y_i is binary with values 0 and 1, the mean $\pi_i = E(y_i)$ is alternatively the probability or proportion $\text{pr}(y_i = 1)$.

A GLM then takes the form

$$g(\pi_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

for some choice of link function g. Common choices of g are the **logit** $g(\pi) = \log[\pi/(1-\pi)]$ and **probit** $g(\pi) = \Phi^{-1}(\pi)$, both of which map $[0, 1]$ to the whole real line.

Logit link allows parameters β_1, β_2, etc., to be interpreted as logarithms of odds ratios. Probit derives from a notional linear regression with normally distributed errors. In practice logit and probit usually give qualitatively similar results.

Directly **linear** probability models, for example

$$\pi_i = \beta_0 + \beta_1 x_i + \beta_2 x_i$$

are problematic when the right side becomes < 0 or > 1.

However, linear models for probabilities (proportions) should not be dismissed out of hand. For some applications they are very natural.

As an example, suppose π_i is the proportion voting Conversative in district i, with x_i and z_i respectively the % votes for Conversative and Fiscist parties at the previous election. The model

$$\pi_i = x_i/100 + \beta z_i/100$$

has the Conversatives gaining a fraction β of the previously-Fiscist vote.

Response type: nominal categories

Write π_{ij} for the probability that the ith respondent is in category j ($j = 1, \ldots, k$). Aim to model the dependence of $(\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})$ on x_i, z_i, etc.

Some possibilities:

- **'multinomial logit' model**: separate linear predictors for each of the logits $\log(\pi_{ij}/\pi_{i1})$ $(j = 2, \ldots, k)$. The choice of 'reference' category is arbitrary, it does not affect the model.
- **'nested logit' model**: e.g., if $k = 4$, find separate linear predictors for $\text{logit}(\pi_{i1})$, $\text{logit}[\pi_{i4}/(1 - \pi_{i1})]$ and $\text{logit}[\pi_{i2}/(\pi_{i2} + \pi_{i3})]$. The particular logits chosen will depend on the context (the research questions of interest). Maybe fewer than $k - 1$ will be needed.

Response type: ordered categories

Methods for nominal categories fail to exploit the extra information in the ordering.

Some better approaches:

- **'cumulative link' models**: with $\gamma_{ij} = \pi_{i1} + \ldots + \pi_{ij}$ $(j = 1, \ldots, k - 1)$, model

$$g(\gamma_{ij}) = \theta_j - \beta_1 x_i - \beta_2 z_i$$

  'Ordered logit' (aka 'proportional odds') and 'ordered probit' models are examples.
  - invariant under merging of categories
  - assumes that the dependence is the same at every 'cut' of the response scale.

Response type: ordered categories (continued)

- **use category scores**: attach a numeric score s_j to each category (e.g., $s_j = j$), and let $y_{ij} = s_j$ when respondent i is in category j. Then construct a linear or generalized linear model for the category scores, treating them as quantitative responses.
  - scores are arbitrary. Different scores may yield different conclusions. Check sensitivity to choice of scores.
  - advantage is availability of linear models and familiar summaries/diagnostics.

In practice this will always be a sensible first analysis (and, depending on the results, may be conclusive).

Response type: duration data

Response y_i is time to an event (e.g., time to death, to employment...)

Aim is to describe the dependence of the distribution of y_i on explanatory variables.

A complication is *censoring*: y_i is not observed, but $y_i > s_i$.

Models for $E(y_i)$ do not deal easily with censoring.

Usually better to model the *hazard* (or 'force of mortality'), $h_i(t) = f_i(t)/[1 - F_i(t)]$. This usually depends on t, though.

The **proportional hazards model** assumes that

$$h_i(t) = h_0(t)\exp(\beta_1 x + \beta_2 z)$$

$$h_i(t) = h_0(t)\exp(\beta_1 x + \beta_2 z)$$

Baseline hazard h_0 may have some assumed form (e.g., Weibull-type monotone hazard, $\gamma t^{\gamma-1}$), or be estimated from the data. It is usually of secondary interest.

The key feature is that effects are multiplicative on the hazard. This is testable, and convenient for interpretation.

Discrete duration data can be treated as an ordered categorical response. The cumulative-link model with *complementary log-log* link function, i.e.,

$$\log[-\log(1 - \gamma_{ij})] = \theta_j - \beta_1 x - \beta_2 z,$$

for the probability of surviving beyond period j, is the discrete-time version of the proportional hazards model.

Random Effects

Nested: sample units $\{ij\}$ are within groups $\{i\}$ (e.g., pupils within schools).

Model intercept and/or coefficients may depend on i, e.g.

$$g[E(y_{ij})] = \beta_{0i} + \beta_1 x_i + \beta_2 z_{ij}.$$

If there are many schools, we would then estimate many intercepts. Sometimes this may be the aim, but more commonly the sampled schools merely represent a larger population (of schools).

A more economical model is then

$$g[E(y_{ij})] = (\beta_0 + b_i) + \beta_1 x_i + \beta_2 z_{ij}.$$

with, say, $b_i \sim N(0, \sigma_b^2)$. Just two parameters describe the intercepts.

Random Effects

Crossed: the grouping involves a cross-classification. For example, pupils j within (classes $c \times$ teachers t) within schools s; or pupils j within (schools $s \times$ home neighborhoods h).

For example, model with random intercepts determined by additive school and neighborhood effects:

$$g[E(y_{shj})] = (\beta_0 + b_s + c_h) + \beta_1 x_i + \beta_z x_{ij}.$$

with, say, $b_s \sim N(0, \sigma_b^2)$ and $c_h \sim N(0, \sigma_c^2)$ independently.

Random Effects

Random slopes: dependence on grouping is not restricted to intercepts.

For example,

$$g[E(y_{ij})] = (\beta_0 + b_i) + \beta_1 x_i + (\beta_2 + c_i) z_{ij}.$$

allows the effect of pupil-specific z_{ij} to depend on school (i).

Care is needed in specifying the joint distribution of the random effects b_i, c_i. Most often the same model should hold regardless of the choice of origin for z_{ij} (e.g., it should not matter whether we use ‘age since birth’ or ‘years since age 4’) — in which case it makes no sense to assume that b_i and c_i are uncorrelated, since that can only be the case for one choice of origin. So here three parameters are needed to describe the distribution of (b_i, c_i): they are σ_b^2, σ_c^2 and ρ_{bc}.

Random Effects

Nonlinear: for example, in repeated responses on the scale

very left	left	neutral	right	very right

subjects may have both a leaning (to left or right) and/or a tendency to prefer the extreme categories (or not).

So if items i with attributes x_i, z_i are presented to subjects s, a suitable random-effects formulation of the cumulative logit model might be

$$\mathrm{logit}(y_{isj}) = c_s(\theta_j + a_s) - \beta_1 x_i - \beta_2 z_i.$$

Threshold parameters θ_j are shifted and squeezed ($c_s < 1$) or spread ($c_s > 1$) for each subject, to reflect their personal interpretation of the response scale.

Here a_s and c_s might be assumed independent, with $E(a_s) = 0$ and $E(c_s) = 1$.

Random or Fixed?

In a random effects model, we replace (say) school-specific parameters with a description of the population of schools.

If there are, say, 15 schools in the study, and lots of information on each school-specific parameter, a fixed-effects model will be best for most purposes. Fifteen is not a large enough sample to allow accurate estimation of population variance.

Random effects models are most effective when the number of groups is large and there is relatively little information per group.

Use of random effects models where the groups represent unique cases of interest (e.g., 15 European nations), seems fundamentally misguided.

Inter-dependent responses

Standard generalized linear models assume responses to be independent, given the linear predictor.

Common situations where this assumption is inadequate:

- units are matched pairs
- units are in clusters, or clusters of clusters...
- repeated measurements (panel data, longitudinal)
- responses close in time
- responses close in space
- multivariate response

Inter-dependent responses

Some general considerations:

- failure of the independence assumption should not be neglected. It will most often result in spurious apparent precision (spurious significance of effects, standard errors misleadingly small, etc.)
- testing for independence is misguided, if the design implies non-independence. Failure to find 'significant' error correlation is *failure*; it is not evidence of independence.
- detailed modelling of the dependence is often unnecessary.
  - sometimes analysis of one or more derived summaries suffices
  - sometimes 'marginal models' suffice

Derived summaries

e.g., repeated measurements $y_{i1}, \ldots, y_{ik}$ (where k is often small, between 2 and 10 say):

level summary: $y_i^* = k^{-1}(y_{i1} + \ldots + y_{ik})$
trend summary: y_i^{**} = regression coefficient of $\{y_{ij} : j = 1, \ldots, k\}$ on time.

Modelling one or other of these may suffice to answer the research question(s) of interest.

With ordered pair data (e.g., before/after measurements), the 'trend' summary would usually be $y_{i2} - y_{i1}$ or y_{i2}/y_{i1}, depending on the nature of y and the questions of interest.

Marginal regression

e.g., clustered data, individuals j within clusters i.

An alternative to random-intercepts model

$$g[E(Y_{ij})] = (\beta_0 + b_i) + \beta_1 x_{ij} + \beta_2 z_{ij},$$

where the Y_{ij} are assumed independent given b_i, is a **marginal model**

$$g[E(Y_{ij})] = \beta_0 + \beta_1 x_{ij} + \beta_2 z_{ij},$$

in which the Y_{ij} are allowed to be correlated within groups.

When g is the identity link (i.e., no link transformation), these are equivalent. More generally they are not: the coefficients then have different interpretations.

Inference in marginal models is handled by the method of 'generalized estimating equations'.

Specialized Models

All applications are specialized, some are more specialized than others...

The models introduced so far should be considered as generic templates for statistical thinking, rather than 'off the shelf' solutions for particular research problems.

In some applications, even the templates are specialized.

Examples include:

- special log-linear models (quasi independence, quasi symmetry) for the analysis of square contingency tables (e.g., social mobility, rater agreement, pair comparisons)
- 'log-multiplicative' interaction structure to measure the strength of association between categorical variables.

e.g., log-multiplicative ‘UNIDIFF’ model for measurement of social fluidity,

$$\log[E(y_{odt})] = \alpha_{ot} + \beta_{dt} + \gamma_t \delta_{od}$$

where (o, d) indexes origin and destination class, and t indexes a set of mobility tables, for example t is country or time-period.

Here the three-way (o, d, t) interaction is structured in a particular way: a common pattern of origin-destination association is assumed to be amplified ($\gamma_t > 1$) or diminished ($\gamma_t < 1$) in the different tables.

When the UNIDIFF simplification holds, it provides a convenient measurement scale for comparing social fluidity across time or place.

Part II

Inference

Part II: Inference

General Notions

Likelihood

Estimating Equations

Simulation, Bootstrap

Inference and Prediction

Simplistically: inference is about parameters (α, β, θ, etc.) whereas prediction concerns unobserved values of variables (not necessarily ‘future’ values).

The distinction is not a clear-cut one. For example, consider the mean of y in a sampled finite population.

The parameters in a model typically refer not to unknown attributes of any particular population, but attributes of the notional *process* (or ‘superpopulation’) that is supposed to have produced the population.

This interpretation makes inference (statistical testing of hypotheses, expression of uncertainty, etc) meaningful even if we have the whole of a finite population as our ‘sample’. Inference concerns the process; all else is prediction.

Models and Correctness

‘Correctness’ of a model is not a practical criterion (almost all useful models are approximations), but is a useful notion for arguments like ‘If this model is correct, then...’

Even notionally, there can be no single correct model.

Recall that a model is a family of distributions. The model is correct if the family includes the distribution from which the data are generated.

Thus, if a model is correct, any larger family containing that model is also correct.

Redundant Parameters Are Bad

The addition of extra terms to a ‘correct’ model will not damage its correctness.

The addition of extra terms will usually improve the degree of fit to the data (as judged by R^2, residual sum of squares, deviance, chi-squared or whatever). But any such ‘improvement’ from redundant terms is **over-fitting**.

Between-model inference (model assessment/selection) aims to avoid over-fitting.

In a model with redundant parameters, information on the quantities of real interest can be seriously diluted by the need to estimate an unnecessarily large number of unknowns. Less obviously, estimation bias can also be introduced.

Within-model Inference

Within a model, the unknown parameters are the objects of interest.

The aim of inference is to identify plausible regions of the model's parameter space, i.e., regions that are compatible with the data.

In practice, except when the dimension of the parameter space is very small, this often reduces to a list of point estimates and an estimated variance-covariance matrix (from which standard errors are extracted).

Likelihood

Write θ for a model's parameter(s).

In essence, the likelihood $L(\theta)$ is the model-determined probability of obtaining the observed data values, when θ is the parameter value. The likelihood is a function of θ, and allows comparison of different candidate values of θ.

More plausible values of θ are those with higher likelihood.

(This notion is due to R. A. Fisher.)

[Actually, the above cannot be used as a definition in the case of continuous data: in that case every single dataset has probability zero! Instead, *density* is used in place of probability, and the same notion of plausibility applies.]

Maximum Likelihood

The maximum likelihood 'principle': as an estimate of θ, compute the value which maximizes $L(\theta)$.

This procedure does not *necessarily* deliver a good estimate. Problems especially when the likelihood has multiple maxima, or the number of parameters is large relative to the amount of data.

But in many simple situations where there is a 'obvious' estimate, the MLE delivers it.

The method's main appeal is

- applicability to an enormous range of (complex) models
- well-developed approximate theory for large samples.

MLE: Large-sample Distribution Theory

Write $\hat{\theta}$ for the MLE of (vector) parameter θ.

Under standard limiting conditions (increasing information per parameter), the MLE is approximately multivariate normal:

$$\hat{\theta} \sim N(\theta, I^{-1}(\theta))$$

where $I(\theta)$ is the *Fisher information matrix* for the model.

Thus estimated standard errors for $\hat{\theta}$ are computed as square roots of diagonal elements of $I^{-1}(\hat{\theta})$ (or, slightly better, of $i^{-1}(\hat{\theta})$, where i denotes the *observed information matrix*).

The main point is that this general method exists, and applies to the vast majority of models used in practice. It is implemented in many computer programs.

Model Comparison

Consider models M_0 and M_1, where M_1 includes M_0 (i.e., M_1 has some additional parameters).

The **generalized likelihood ratio test**: if M_0 is correct, then in large samples the 'twice log likelihood ratio' statistic

$$W = 2(\hat{l}_1 - \hat{l}_0)$$

has approximately the chi-squared distribution with degrees of freedom equal to the number of additional parameters.

(This result due to S. S. Wilks)

This provides a scale on which to judge whether the additional parameters are worthwhile, or redundant: if the value of W is unusually large relative to the reference chi-squared distribution, that is evidence of non-redundancy.

Bayesian Inference

An alternative use of the likelihood: rather than maximize $L(\theta)$, take the average of $L(\theta)$, weighted by a specified **prior distribution** $p(\theta)$, and normalized so that the result is a probability distribution for θ.

This produces a Bayesian **posterior distribution**, $p(\theta|y)$.

Inference on θ is then carried out by making probability statements derived from $p(\theta|y)$.

This avoids large-sample approximations. It has the drawback that the results can depend strongly on the specified prior $p(\theta)$. Computation can also be problematic, involving high-dimensional integrals which are typically approximated by McMC simulations.

AIC, BIC, etc.

The various ‘information criteria’ (AIC, BIC, etc.) aim to automate the balance between

- fit to the data (model has high likelihood value)
- exclusion of redundant parameters

The typical form of such criteria is

$$\text{*IC} = l - \text{penalty}(n, p),$$

which *penalizes* the model likelihood for the number of parameters (p) in relation to the data size (n).

These criteria can be helpful for screening when the set of candidate models is very large, but are no substitute for careful model criticism. The AIC criterion seems particularly well suited to automated prediction problems.

Estimating Equations

Most of the standard estimation methods are based on the solution of a system of **unbiased estimating equations** (one equation for each unknown parameter). For example, least squares in the linear model solves

$$X^T(y - X\beta) = 0$$

while maximum likelihood for model parameters θ solves

$$\nabla l(\theta) = 0$$

Unbiased means that the expectation of left and right sides of the equations are equal.

Subject to the some technical conditions, estimates obtained as solutions to unbiased estimating equations are consistent, that is they tend in probability to the true value of β or θ.

Quasi-likelihood

When y has known variance-covariance matrix V, generalized least squares solves

$$D^T V^{-1}[y - \mu(\beta)] = 0$$

where D denotes the partial-derivatives matrix $\partial\mu/\partial\beta$.

Quasi-likelihood: when y has variance-covariance matrix $V(\beta)$, estimate β by solving

$$D^T V(\beta)^{-1}[y - \mu(\beta)] = 0$$

These are unbiased estimating equations, even if the assumed form of V is incorrect.

Applications include: GLMs for over-dispersed count data; marginal GLMs for panel data (the method of ‘generalized estimating equations’).

Quasi-likelihood: Inference

Inference is based on large-sample normality, and an estimated variance-covariance matrix.

Two styles of estimated vcov matrix:

'model-based': assumes that V has the correct form. The variance-covariance matrix is then $(D^T V^{-1} D)^{-1}$, essentially the inverse of the 'quasi information' matrix.

'robust': allows that the form of V may be wrong. The large-sample theory then gives the 'information sandwich' formula

$$(D^T V^{-1} D)^{-1} [D^T V^{-1} \operatorname{cov}(y) V^{-1} D] (D^T V^{-1} D)^{-1}$$

In practice, 'robust' needs very large n in order to be reliable.

Over-dispersed Counts

e.g„ Poisson-type counts but with $\operatorname{var}(y_i) > E(y_i)$; binomial-type counts but with $\operatorname{var}(y_i) > m_i \pi_i (1 - \pi_i)$.

Overdispersion can be due to missing covariates, clustering.

If the 'standard' variance function is $V(\mu)$ (e.g., Poisson has $V(\mu) = \mu$), an overdispersed version is $\phi V(\mu)$ with $\phi > 1$.

In this case the quasi-likelihood equations are identical to those for maximum likelihood: the estimates are unchanged.

The 'model-based' vcov matrix is $\phi(D^T V^{-1} D)^{-1}$, which is as for MLE but inflated by factor ϕ.

So the only change needed is to inflate all standard errors by the square root of an estimate of ϕ.

GEE for Marginal Models

The 'generalized estimating equations' method for marginal panel-data models is

- specify a 'working' form for the V matrix, which allows some guessed form of correlation among an individual's repeated measurements
- solve the quasi-likelihood equations using the 'working' V for $\hat{\beta}$
- use the 'information sandwich' variance-covariance estimate, to protect against mis-specification of V.

The simplest working correlation matrix is the identity matrix, corresponding to no correlation among the repeated measurements. This is likely to be inefficient in many applications, but it is often the default choice.

Simulation, Bootstrap

Computer simulation is an alternative to reliance on large-sample theory.

Two main approaches:

model-based: generate new sample responses y from the fitted model, and re-analyse by the same method

nonparametric: generate a whole new dataset by resampling from the original data, with replacement, and re-analyse by the same method.

With either approach, repeated application (say, 1000 times) allows assessment of the repeated-sampling properties of the method used.

This can demand substantial computer time.

Part III

Diagnostics

General Principles

Some common points are:

- all estimates and standard errors are based on **modelling assumptions**, which need to be checked
- key quantities are **residuals**, designed to exhibit **no structure** when the model is correct
- chi-squared and other such **lack-of-fit statistics** summarize the residuals crudely, and are not enough on their own
- the **influence** of particular data items on different aspects of the model might also be considered.

Residuals

In a linear model, residuals are $y_i - \hat{\mu}_i$.

More generally, the form of residuals may differ (Pearson residuals, deviance residuals, etc.). The aim in all cases is a set of quantities which exhibit no structure when the model is correct.

There are special residuals for special purposes, e.g., partial residuals for checking an omitted covariate.

'Exhibit no structure' leaves open *any* imaginable way of looking at a model's residuals!

There are, however, some standard devices: scatterplots and summary tables.

Some standard residual plots:

- residuals versus fitted values. Examine vertical slices for homogeneity. Can reveal un-modelled curvature, or un-modelled mean-variance relationship.
- residuals versus x-variables and candidate x-variables.
- sorted residuals versus the quantiles or expected order statistics for a sample of the same size from a standard distribution (e.g., $N(0, 1)$, a 'normal Q-Q plot'). This is less a check on distributional shape — in many models we do not expect normally-distributed residuals, for example — than a screening device for outliers.

When the predictor variables are categorical, *tabulation* of averaged residuals can give useful clues to model inadequacies.

Lack-of-fit Tests

A lack-of-fit test typically is based on a statistic which summarizes the residuals (e.g., sum of squared Pearson residuals is the standard chi-squared statistic for count-data models).

Significant lack of fit? Look at the residuals to see why/where.

Non-significant? The summary may not reveal important patterns present in the residuals. (A different summary might have.) Look at the residuals, or risk getting things wrong.

With very large datasets, lack-of-fit tests can indicate significant evidence against the model even under the tiniest of departures. Attention should then focus on whether those departures are substantively important.

Influence

In linear and generalized linear models, the **leverages** (one for each observation in the data) determine the impact that each observation has on the model fit.

Leverage is a function of position in the space of covariate values (x-values): units that are extreme in one or more x variables will exert high leverage on the fit.

The leverages are diagonal elements of the model's 'hat matrix', and are often denoted by h_i. The average value of h_i over all the observations is p/n, where p is the number of parameters in the model and n the data size.

A further set of unit-specific quantities, the **Cook's distances** bring together leverages and residuals, to indicate the combined effect of leverage and deviation from the model.

Part IV

Interpretation

Part IV: Interpretation

Precision and Significance

Parameterization

Conditional Independence, Graphical Models

Causality

Precision and Significance

Some general points:

- any statement about a parameter or combination of parameters should be accompanied by a statement of precision (e.g., a standard error or confidence interval)
- significance testing of evidence against the zero hypothesis for a parameter is done in order to establish whether we really know the direction of an effect
- lack of significance (of evidence against the zero hypothesis for a parameter) does not imply that the parameter's value is small. Lack of significance reflects lack of evidence.
- effect sizes should not be compared by a comparison of significance levels.

Parameterization

Parameterization is simply the way in which a model is represented (its 'coordinate system').

Parameterization does not affect a model's meaning in any way. However, particular parameters in different representations of a model have different meanings.

Arbitrary aspects of parameterization often include:

- the origin and scale of continuous predictor variables.
- the 'reference category' used in coding the levels of a categorical predictor (factor)

Intercept

The intercept (or constant term) in a model is the estimated value of the predictor when all of the explanatory variables are zero. This may be

- an object of no substantive relevance (e.g., if x is respondent's weight in kg — no-one has zero weight!)
- estimated very poorly if 'all-zero' is a long way from the centre of the data

It will often be helpful to change the origin of some or all x-variables to make the intercept meaningful. One possibility is to centre x-variables on their means, i.e., work with $x^* = x - \bar{x}$; but this will not always make sense, e.g., if x is binary.

Standardization

The scale of an x-variable affects the scale of its coefficient: e.g., measurement in *kg* requires a β-value 1000 times as big as does measurement in *g*.

Comparison of effect sizes is made difficult by different scales: for example if x is in *kg* and z is in \$.

It is *sometimes* useful to convert variables to a scale on which 'one unit' means 'one standard deviation', i.e., to re-scale x to x/σ_x, with σ_x the standard deviation of x in the sample (or population, or some other standardizing group). Equivalently, multiply β by σ_x.

Care is needed. Standardization will rarely make sense, for example, in the case of binary x (e.g., 0/1 for male/female), where the unstandardized β has a clear interpretation.

Coding of Categorical Predictors

A standard device in representing the effect of a k-level factor is to use $k - 1$ 'dummy variables'. The coefficients are then interpreted as contrasts with the omitted ('reference') level.

A change of reference level does not change the model: any contrast is easily computed from the coefficients of any representation.

Note, however, that the reported *standard errors* are specific to the choice of reference level. For example, if level 1 is the reference, no standard error is routinely made available for $\hat{\beta}_3 - \hat{\beta}_2$, which requires knowledge of $\mathrm{cov}(\hat{\beta}_2, \hat{\beta}_3)$.

A useful device to overcome this problem is the notion of *quasi-variances*, as implemented in the *R* package *qvcalc*.

Relative Sizes of Effects

The comparison of effect sizes is a controversial topic.

For two x-variables measured in the same units, unstandardized coefficients can be compared directly, via the ratio β_1/β_2. Otherwise, standardization may be needed in order for such a calculation to become remotely meaningful.

The notion of comparing standardized effects can be extended also to comparing the combined effects of two *groups* of terms in the predictor, e.g., 'school effects' and 'neighborhood effects': if $(\beta_1 x_i + \beta_2 z_i + \beta_3 t_i)$ is such a group of terms, its effect is summarized as the variance, either in the sample or in some standardizing population, of $\hat{\beta}_1 x + \hat{\beta}_2 z + \hat{\beta}_3 t$. The *R* package *relimp* provides tools for inference on the comparison of such summaries.

Interaction

Interaction effects always complicate interpretation.

Sometimes interaction effects are inevitable. Occasionally, though, interaction can be re-represented as main effect, by a suitable transformation or re-coding.

As a trivial example, consider two binary variables x (1 if mother eats chips, 0 otherwise) and z (1 if father eats chips, 0 otherwise). A model including the main-effect and interaction terms

$$+x + z - 2xz$$

would be more simply expressed as

$$+w$$

where w is defined as 1 if mother and father differ in their liking for chips, 0 otherwise.

Conditional Independence, Graphical models

Some models, e.g., some log-linear models for contingency tables, can be interpreted in terms of the implied conditional independence relationships among the variables involved.

A useful representation of such models is in terms of a *graph*, in which missing edges indicate conditional independence: hence the models are known as *graphical models*.

Conditional independence often allows seemingly complex interaction structures to be broken down into smaller parts for interpretation.

Conditional independence is also central to the definition of many *latent variable* models, e.g., factor analysis, structural equation models, latent class models, etc.

Example: drug use in high school (Agresti, *Categorical Data Analysis*, p178)

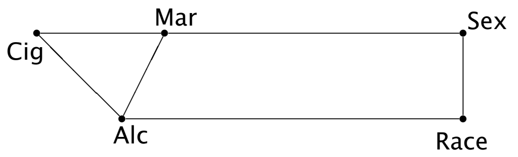
A five-dimensional table: alcohol, cigarette and marijuana use for high school seniors in Dayton Ohio, classified by sex and race.

	Race:	White				Other			
	Sex:	Female		Male		Female		Male	
Alcohol	Cigarette	Marijuana use							
use	use	Yes	No	Yes	No	Yes	No	Yes	No
Yes	Yes	405	268	453	228	23	23	30	19
	No	13	218	28	201	2	19	1	18
No	Yes	1	17	1	17	0	1	1	8
	No	1	117	1	133	0	12	0	17

Analysis of deviance in the usual way leads to detailed consideration of the log-linear model

```
SEX*RACE + SEX*MAR + RACE*ALC +
          MAR*ALC*CIG
```

which has deviance 24.8 on 19 degrees of freedom. This is a graphical model (it is characterized by the absence of some 2-way interactions), whose graph is:



Cigarette use can thus be safely summarized in a table collapsed over sex and race (proportion using cigarettes in each category):

Alcohol	Marijuana use	
use	Yes	No
Yes	95%	54%
No	60%	13%

Note that here the figure of 60% smoking in the ‘marijuana but not alcohol’ cell is based only on 5 cases and should therefore be treated with caution.

Causality

Deliberately left to last!

Except where the data arise from a carefully controlled study (e.g., a randomized experiment), inferring causal relationships from a statistical model alone is unwise.

A statistical model represents *one* possible mechanism by which the data may have been generated. It does not rule out others, in particular mechanisms involving unmeasured variables.

Statistical models do help to suggest possible causal mechanisms, and they are useful also in providing an empirical framework for the refutation (or otherwise) of causal explanations that come from substantive theory.

Part V

Some Things to Worry About

Very briefly, a very partial list:

- do the results make sense?
- are there *mediating variables*, or are there *interactions*, which could qualitatively change the interpretation?
- co-linearity: have I perhaps included the same effect twice under different names?
- the impact of assumptions, and of missing data (missing cases; missing variables). Consider sensitivity analysis, in which modelling assumptions and/or assumptions about missingness are varied.

Some further reading

Cox, D R and Snell, E J (1981). *Applied Statistics: Principles and Examples*. Chapman & Hall.

Cox, D R (1990). Role of Models in Statistical Analysis. *Statistical Science*.

Cox, D R and Wermuth, N (2001). Some Statistical Aspects of Causality. *European Sociological Review*.

Freedman, D A (1991). Statistical Models and Shoe Leather. *Sociological Methodology*.

Snijders, T A B and Bosker, R J (1999). *Multilevel Analysis*. Sage.