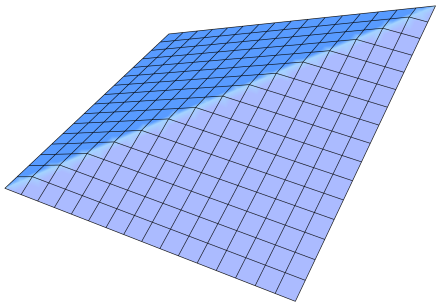

STA 2201H: Methods of Applied Statistics II

Prof. Patrick Brown • Winter 2019 • University of Toronto
W 2:00-5:00PM in BL 325

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Wednesday, January 09, 2019

1. Statistical Practice and Reproducible Communication

Applied Statistical Modelling

Applied statistics uses data to solve problems in which uncertainty and randomness are unavoidable. Its distinguishing task is inference about phenomena that have not been observed directly. **Data science** is broader: it includes the collection, management, processing, analysis, visualization, and interpretation of heterogeneous data. Applied statistics is therefore part of data science, but many data science tasks concern quantities that are observed directly and require no probabilistic inference.

The statistical contribution begins with two principles. First, probability is the appropriate language for uncertainty. A **stochastic model** describes uncertainty in the data, while probabilistic inference describes uncertainty in the conclusions. Second, context matters. What can be learned from a numerical data set depends on how the observations were collected, which variables were measured, and what scientific mechanism generated them. Careful design can prevent uncertainty or bias that no subsequent analysis can repair. These themes are developed further by Diggle's account of statistics as a data science and Breiman's discussion of two modelling cultures.

Statistical computing has repeatedly changed which models can be used in practice. Generalized linear models became widely accessible through software implementing iteratively reweighted least squares. **Markov chain Monte Carlo** methods later made complicated Bayesian models computationally feasible. More recent tools include Hamiltonian Monte Carlo, automatic differentiation, variational approximations, and **integrated nested Laplace approximation**. The common lesson is that software expands the statistician's reach, but does not replace the need to formulate an appropriate stochastic model and interpret it in context.

Breiman distinguished a **data modelling culture**, which specifies a stochastic mechanism and estimates its parameters, from an **algorithmic modelling culture**, which emphasizes predictive performance. The contrast is related to several recurring choices: parametric models versus machine learning, explanation versus prediction, and low bias versus low variance. Experimental design and survey sampling provide a third perspective by controlling how information is collected.

For squared-error prediction, the expected error is decomposed schematically as

$$\text{prediction error} = \text{irreducible noise} + \text{bias}^2 + \text{variance}.$$

Variance is caused by uncertainty in estimated parameters and typically decreases as the sample size grows. **Bias** is caused by systematic model misspecification and need not decrease with sample size. In a modest data set, variance may dominate and a carefully chosen parametric model can be extremely effective. In a very large data set, conventional estimation variance may be negligible, so model bias becomes the limiting error. More data do not correct a systematically inappropriate

model.

Figure 1 records the rise of open source statistical computing. Its purpose is historical rather than inferential: popularity is only a proxy for influence, but the rapid growth of R illustrates how extensible software changed applied statistical practice.

Unless otherwise stated, numerical displays are schematic reconstructions generated from the stated models or reported summaries. They preserve the inferential point but are not reanalyses of the original microdata.

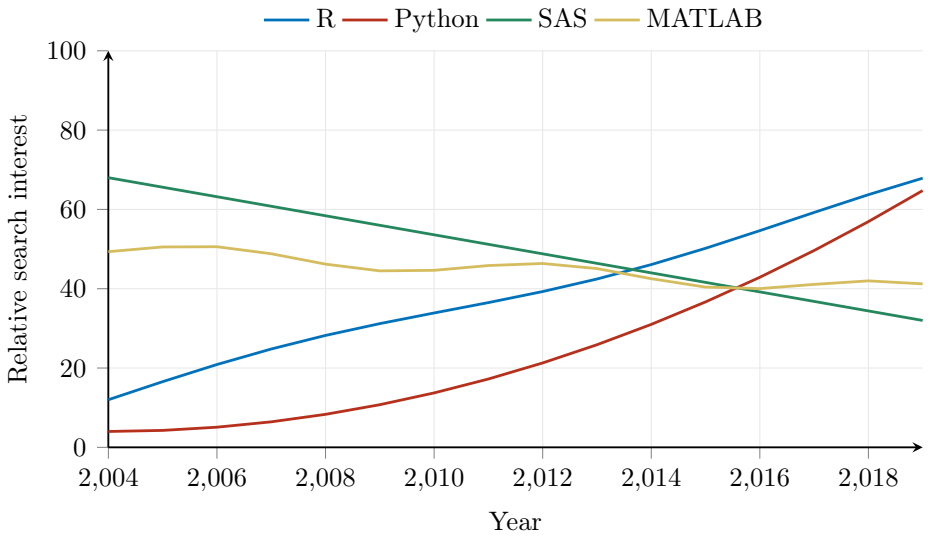

Figure 1

The principal model families for non-Gaussian and dependent data are generalized linear models, mixed and hierarchical models, semiparametric models, survival models, and spatial models. Many of these models contain latent Gaussian variables, for which **Bayesian inference** and integrated nested Laplace approximation are especially useful.

Four motivating studies show why a single modelling template is not enough.

Example 1.1 Each launch has m_i O-rings, of which Y_i are damaged. The scientific question is whether damage is more likely in cold weather after allowing for launch pressure. A Gaussian response model is unsuitable because Y_i is a count bounded by m_i. A natural starting point is

$$Y_i \mid m_i, p_i \sim \text{Bin}(m_i, p_i),$$

with the failure probability p_i modelled as a function of temperature and pressure. Pressure may confound the apparent temperature association, so both variables should be considered.

The observed damaged proportions in the shuttle data are shown in Figure 2. The discreteness and unequal denominators are part of the data generating structure, not imperfections to be smoothed away before modelling.

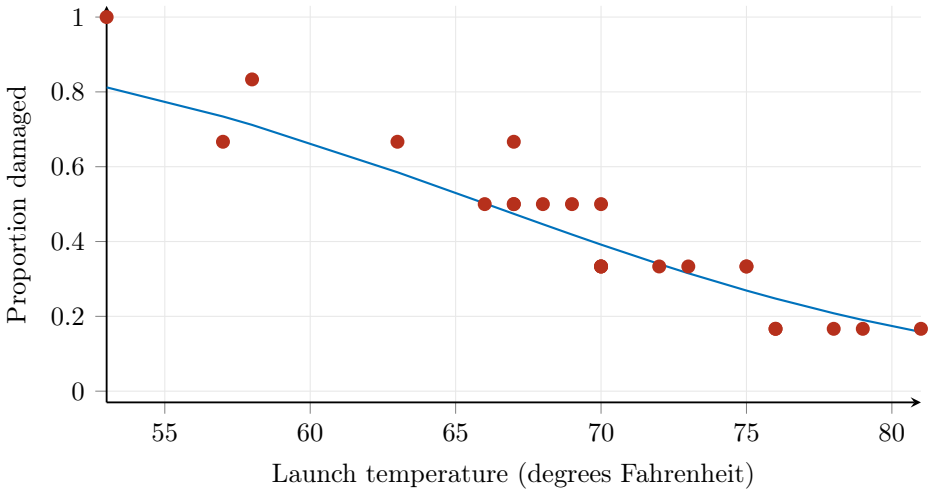

Figure 2

Example 1.2 The 1974 Fiji Fertility Survey recorded the number of children born to each respondent, together with age, age at marriage, residence, literacy, ethnicity, and related variables. One question is whether literate women tend to have smaller families, and whether an observed difference can instead be explained by earlier marriage among women who are not literate. A count model such as a Poisson regression supplies a starting point, but the exposure time and the possibility of extra zeros or overdispersion must also be examined.

The empirical distributions in Figure 3 compare family size by literacy. The plot motivates a model; it does not by itself adjust for age at marriage or other confounding variables.

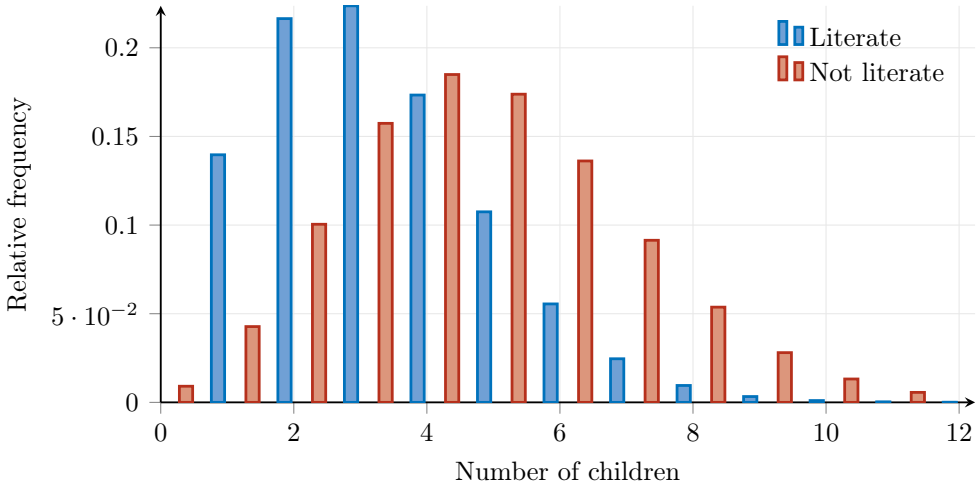

Figure 3

Example 1.3 The 2014 American National Youth Tobacco Survey recorded cigarette, chewing tobacco, snuff, and hookah use together with age, sex, race, rural or urban status, state, school, and other characteristics. Binary regression can describe associations with each form of tobacco use. Because respondents are grouped within schools and states, an adequate model may also require random effects to represent dependence within those clusters.

Example 1.4 Records from 2015 and 2016 classify deaths involving police by race or ethnicity, state, whether the person was armed, and the manner of death. A direct comparison of the proportions killed while unarmed can be misleading because the racial composition of recorded deaths differs sharply among states. Moreover, a database containing deaths but not all police encounters cannot directly estimate the probability of death conditional on an encounter. This distinction between the observed sampling mechanism and the desired probability is central to case-control reasoning.

These examples illustrate a general workflow. Begin with the scientific question and the observation process, choose a response distribution that respects the support of the data, specify a systematic component for relevant explanatory variables, represent dependence when observations share a cluster, time, or location, and report uncertainty on an interpretable scale. The slogan that all models are wrong but some are useful is valuable only when accompanied by a precise account of what the model is intended to answer.

Scientific Writing and Reproducible Analysis

A statistical report must be adapted to its audience. A methodological paper must establish that a proposed method is correct and useful. An application paper should explain the model briefly, interpret its parameters in the language of the application, and connect its conclusions to the scientific question. A consulting report must solve the client's problem without requiring the client to reconstruct the analysis. A thesis or extended report usually contains more background, but it should not display material merely to demonstrate the author's knowledge.

A conventional applied report contains the following components.

1. The **title** identifies the scientific subject and, when useful, the study population, place, or period.
2. The **abstract** summarizes the problem, data, method, principal findings, and conclusion in a single self-contained paragraph.
3. The **introduction** explains the background, states the research question, describes the available data and their origin, and reviews closely related work when the genre requires it.
4. The **methods** section describes the study population, variables, sample size, observation period, exploratory analysis, statistical model, inferential method, model selection rule, and diagnostic strategy.
5. The **results** section first summarizes the data and then presents estimates, uncertainty intervals, tests, and figures from the final model.
6. The **discussion** interprets the results in context, answers the research questions, identifies important limitations, and states the conclusions without overstating causality.
7. The **references** contain every cited source in one consistent format. An appendix is optional and is reserved for material such as detailed derivations, code required for reproducibility, or complementary analyses.

The methods section should explain both what was done and why the method is appropriate. Standard methods require little general justification, but the analysis must still connect the model to the data. For example, repeated measurements on the same animal are likely to be related, as are observations on animals from the same farm, so a multilevel model with farm and animal random effects may be justified. The model selection and assessment rules should also be stated before the reader sees their results.

The results section should emphasize the selected analysis rather than presenting every model that

was tried. Parameter estimates should be transformed when the natural model scale is difficult to interpret. For example, a logistic regression coefficient is usually more informative as an odds ratio, and a log linear coefficient is usually more informative as a rate ratio. Competing models should be shown only when the comparison answers a substantive question, demonstrates sensitivity, or documents an important diagnostic decision.

The discussion returns to the scientific scale. Statements such as "the random effect variance was 1.342" are incomplete unless the magnitude is compared with an effect or with relevant variation in the response. Statistical significance is not a substitute for scientific importance, and association should not be presented as causation unless the design and assumptions justify that interpretation.

References, results, code, and figures should be generated in a **reproducible workflow**. A useful separation is

source data and code ⟶ derived numerical and graphical results ⟶ typeset report.

The analysis code should be able to regenerate every reported quantity, while the report source should refer to those generated results consistently. Long computations can be cached, but the cache is a convenience rather than a replacement for retaining the code and inputs that created it.

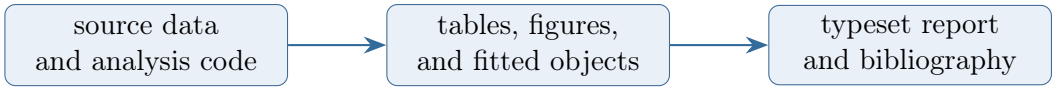

Figure 4

The three stages of the reproducible workflow are summarized in Figure 4.

LaTeX supports structural commands, automatic cross references, mathematical typesetting, floating tables and figures, and automatically formatted bibliographies. Markdown provides a lighter source format and can be converted to LaTeX, HyperText Markup Language, and other document formats. **Literate programming** tools such as knitr evaluate R code embedded in a source document and insert its output into the typeset result.

Only code that clarifies or reproduces the analysis belongs in the main exposition. A short computational fragment might be retained as follows.

```
fit <- glm(damaged / total ~ temperature + pressure,
           weights = total, family = binomial, data = shuttle)
confint(fit)
```

Routine import, package installation, printing, and formatting commands should be kept in the reproducibility files rather than interrupting the mathematical discussion. Code that takes a long time to run may save a fitted object and reload it when the source data and analysis specification have not changed. The final report must nevertheless make clear how every result was obtained.

Lecture 2 — Wednesday, January 16, 2019

2. Generalized Linear Models and Likelihood Inference

The Generalized Linear Model Framework

Definition 2.1 A **generalized linear model** has independent responses $Y_1, \ldots, Y_n$ from a specified exponential family, satisfying

$$Y_i \sim G(\mu_i, \boldsymbol{\theta}), \qquad \mathbb{E}[Y_i] = \mu_i, \qquad \text{and} \qquad h(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

Here, G is the response family, μ_i is its mean, $\boldsymbol{\theta}$ contains any additional common nuisance or dispersion parameters, h is the **link function**, $\boldsymbol{x}_i$ is the vector of explanatory variables for observation i, and $\boldsymbol{\beta}$ is the vector of regression coefficients.

The response distribution, **linear predictor**, and link play different roles. The distribution determines which values the response may take and how its variance changes with its mean. The linear predictor $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ describes systematic variation. The link converts the response scale parameter into a quantity that may vary over the whole real line.

Ordinary least squares is the Gaussian special case

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad \text{and} \qquad \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

so G is Gaussian, $\boldsymbol{\theta} = \sigma^2$, and h is the identity link.

Definition 2.2 A random variable Y has a **Poisson distribution** with mean $\lambda > 0$ if

$$\mathbb{P}(Y = y) = \frac{\lambda^y \exp(-\lambda)}{y!}, \quad y = 0, 1, 2, \ldots.$$

Its mean and variance are both λ. If Y_1 and Y_2 are independent Poisson variables with means λ_1 and λ_2, then

$$Y_1 + Y_2 \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

The two probability mass functions in Figure 5 show how the Poisson distribution changes as its mean increases.

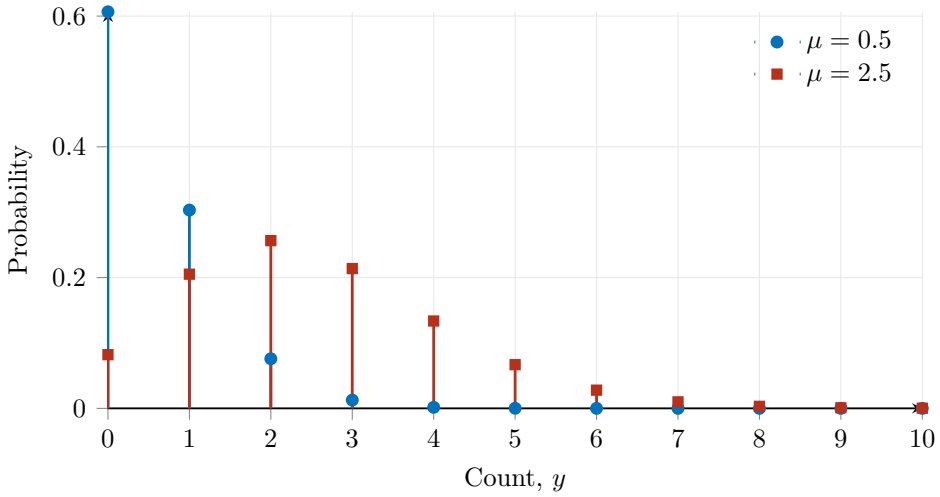

Figure 5

Poisson regression uses

$$Y_i \sim \text{Poisson}(\mu_i) \qquad \text{and} \qquad \log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

The logarithmic link ensures that every fitted mean is positive and makes multiplicative changes in the mean additive on the linear predictor scale.

Definition 2.3 A distribution is **infinitely divisible** if, for every positive integer K, a variable with that distribution can be represented as the sum of K independent and identically distributed variables.

Gaussian, Poisson, gamma, and lognormal distributions are infinitely divisible. For example, if

$$Y_i = \sum_{k=1}^{K} X_{ik} \qquad \text{and} \qquad X_{ik} \sim \text{Poisson}(\lambda_i/K)$$

independently, then $Y_i \sim \text{Poisson}(\lambda_i)$. This property provides a useful modelling argument when an observed total is the accumulation of many smaller, unobserved contributions. A monthly cigarette count may be regarded as the sum of daily counts, for example. The lognormal result is nontrivial and was proved by Thorin (1977); its convolution roots are not generally lognormal. Uniform distributions are not infinitely divisible under addition. A binomial distribution with a fixed number of trials can be divided only when the number of trials is compatible with the proposed subdivision. These observations guide model construction, but they do not imply that every response must be an accumulated sum.

For a binomial response, the standard model is

$$Y_i \sim \text{Bin}(N_i, \mu_i) \qquad \text{and} \qquad \log\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where μ_i is the success probability. When $N_i = 1$, the response is Bernoulli. The **logit link**

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

maps a probability in $(0, 1)$ to the real line.

Example 2.4 Let Y_i be the number of children born to woman i, let O_i be the number of months from her first marriage to the survey date, and let μ_i be her birth rate per month. The model may be written in either of the equivalent forms

$$Y_i \sim \text{Poisson}(O_i\mu_i) \qquad \text{and} \qquad \log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

or

$$Y_i \sim \text{Poisson}(\rho_i) \qquad \text{and} \qquad \log(\rho_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(O_i).$$

The known term $\log(O_i)$ is an **offset**: it has coefficient one and accounts for unequal exposure. The first form communicates the rate directly, while the second is the form supplied to regression software.

Definition 2.5 A random, locally finite collection of event times on $[0, \infty)$ is a **homogeneous Poisson process** with intensity $\lambda > 0$ if the following conditions hold.

1. For every bounded time interval A, the number $N(A)$ of events in A has distribution Poisson$(\lambda|A|)$.
2. For every finite collection of pairwise disjoint bounded intervals, the corresponding counts are jointly independent.

Conditional on $N(A) = n$, the unordered locations in a bounded interval A have the same distribution as n independent uniform points on A; equivalently, the chronologically ordered event times are their order statistics. In Example 2.4, the interval A_i has length O_i, so a homogeneous process with rate μ_i gives $Y_i \sim \text{Poisson}(\mu_i O_i)$. The model respects the nonnegative, right-skewed count response and gives μ_i an interpretation independent of exposure. It is nevertheless only a starting point. Births cannot occur independently at arbitrarily short intervals, fertility changes with age, and unobserved differences in fertility and desired family size may produce overdispersion.

Likelihood Based Estimation and Testing

Under the independence assumption in Definition 2.1, the **likelihood function** is the joint density regarded as a function of the unknown parameters. The likelihood and its logarithm are

$$\pi(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f_G(Y_i; \mu_i, \boldsymbol{\theta})$$

and

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log f_G(y_i; \mu_i, \boldsymbol{\theta}).$$

The maximum likelihood estimators satisfy

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\beta}, \boldsymbol{\theta}} \ell(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y}).$$

They are the parameter values under which the observed data have the greatest likelihood.

The **score** for coefficient β_p follows from the chain rule:

$$\frac{\partial \ell}{\partial \beta_p} = \sum_{i=1}^{n} \left. \frac{\partial \log f_G(y_i; \mu, \boldsymbol{\theta})}{\partial \mu} \right|_{\mu = h^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})} \left. \frac{\partial h^{-1}(\eta)}{\partial \eta} \right|_{\eta = \boldsymbol{x}_i^\top \boldsymbol{\beta}} x_{ip}.$$

For standard generalized linear models, both derivatives have explicit forms. Newton–Raphson or Fisher scoring therefore finds the maximum quickly. When G belongs to the exponential family, the computation can be expressed as iteratively reweighted least squares.

Let $\mathbf{H}(\boldsymbol{\beta})$ denote the **Hessian** of ℓ with respect to $\boldsymbol{\beta}$. A second order Taylor expansion about $\boldsymbol{\beta}_0$ gives

$$\ell(\boldsymbol{\beta}) \approx \ell(\boldsymbol{\beta}_0) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \nabla \ell(\boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^\top \mathbf{H}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

At the maximum, $\nabla \ell(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. The local likelihood is therefore approximately Gaussian, with covariance given by the inverse information:

$$\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\left(\boldsymbol{\beta}, \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1}\right) \qquad \text{and} \qquad \mathbf{I}(\widehat{\boldsymbol{\beta}}) = -\mathbf{H}(\widehat{\boldsymbol{\beta}}).$$

The square roots of the diagonal entries of $\mathbf{I}^{-1}$ are the usual information based standard errors.

It is important to distinguish the statistical model, the inferential framework, and the numerical algorithm. Generalized linear models, time series models, and sampling models specify probability structures. Likelihood based procedures, Bayesian procedures, method of moments procedures, and partial likelihood procedures are inferential frameworks. Least squares, iteratively reweighted least squares, Markov chain Monte Carlo, integrated nested Laplace approximation, and the lasso are algorithms or computational procedures.

Theorem 2.6 (Likelihood ratio theorem) Suppose that the regularity conditions for likelihood inference hold. Let the null hypothesis constrain $\beta_k = C_k$ for each $k \in \Omega$, where Ω contains r indices. If $\widehat{\boldsymbol{\beta}}$ is the unrestricted maximum likelihood estimator and $\widehat{\boldsymbol{\beta}}_C$ is the constrained estimator, then, under the null hypothesis,

$$2\left\{\ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}) - \ell(\widehat{\boldsymbol{\beta}}_C; \boldsymbol{y})\right\} \xrightarrow{d} \chi^2_r.$$

The null model must be nested in the alternative model: every distribution allowed under the null hypothesis must also be allowed under the alternative. For Poisson and binomial models, there is no additional dispersion parameter. In other response families, a nuisance parameter may factor out or may require separate numerical optimization. Software often estimates such a parameter and then treats it as fixed when reporting approximate inference for $\boldsymbol{\beta}$; this approximation deserves attention when the nuisance parameter is poorly determined.

Interpretation and Profile Likelihood

Example 2.7 The following fit uses the number of damaged and undamaged rings as a grouped binomial response. Temperature is the parameter of scientific interest, while pressure is retained as a possible confounder.

```
shuttle$undamaged <- shuttle$m - shuttle$r
response <- with(shuttle, cbind(r, undamaged))
fit <- glm(response ~ temperature + pressure,
           family = binomial, data = shuttle)
```

The fitted linear predictor is

$$\widehat{\text{logit}}(\mu_i) = 2.5202 - 0.09830\,\text{temperature}_i + 0.008484\,\text{pressure}_i.$$

The maximized log likelihood is -15.05294.

The information based estimates are

parameter	estimate	standard error	p-value
intercept	2.520	3.487	0.470
temperature	−0.098	0.045	0.029
pressure	0.008	0.008	0.269

The likelihood ratio statistic comparing this model with an intercept only model is 7.6847 on two degrees of freedom, with $p = 0.02144$. Thus, temperature and pressure are jointly associated with ring damage.

For a logistic model with

$$\text{logit}(\mu_i) = \sum_{p=1}^{P} x_{ip}\beta_p,$$

suppose two observations agree in every covariate except that $x_{2q} = x_{1q} + 1$. Their fitted odds

satisfy

$$\beta_q = \log\left(\frac{\mu_2}{1-\mu_2}\right) - \log\left(\frac{\mu_1}{1-\mu_1}\right)$$

and hence

$$\exp(\beta_q) = \frac{\mu_2/(1-\mu_2)}{\mu_1/(1-\mu_1)}.$$

Thus, β_q is a log odds ratio and $\exp(\beta_q)$ is an **odds ratio**. The intercept gives the baseline log odds when all other covariates equal zero. If zero lies far outside the observed range, the intercept should be made meaningful by centring the covariates.

For the shuttle data, centring temperature at 70 degrees Fahrenheit and pressure at 200 units gives a baseline fitted probability of 0.065, with an approximate 95% interval from 0.027 to 0.149. Each one degree increase in temperature multiplies the odds of damage by 0.906, with interval $(0.830, 0.990)$, corresponding to an estimated 9.36% decrease in the odds. Comparisons over an interquartile range are often more interpretable when covariates have different units. The interquartile odds ratios are 0.455 for an eight degree increase in temperature and 2.888 for a 125 unit increase in pressure; the corresponding intervals are $(0.225, 0.921)$ and $(0.440, 18.942)$.

Definition 2.8 Partition a parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, where $\boldsymbol{\psi}$ is the parameter of interest and $\boldsymbol{\lambda}$ is a nuisance parameter. The **profile likelihood** for $\boldsymbol{\psi}$ is

$$\widetilde{L}(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda}} L(\boldsymbol{\psi}, \boldsymbol{\lambda}).$$

For the shuttle model, we may profile over the intercept and retain the temperature and pressure coefficients. The left panel of Figure 6 shows the resulting two dimensional profile deviance; its dashed boundary is the 95% likelihood ratio cutoff $\chi^2_{2,0.95} = 5.99$. The right panel profiles over both the intercept and pressure coefficients to leave a one dimensional deviance for temperature, where the horizontal line is the cutoff $\chi^2_{1,0.95} = 3.84$.

Profile likelihood can reveal asymmetry or parameter dependence that the quadratic information approximation conceals. Information based intervals are nevertheless common because they are fast and usually adequate in regular, well identified problems.

The logit link is not the only binomial link with a process interpretation. Suppose that failures for ring j on launch i follow a Poisson process with rate λ_i during a mission of duration δ, where

$$\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

If $K_{ij} \sim \text{Poisson}(\delta\lambda_i)$ is the number of latent failures and Z_{ij} records whether at least one failure

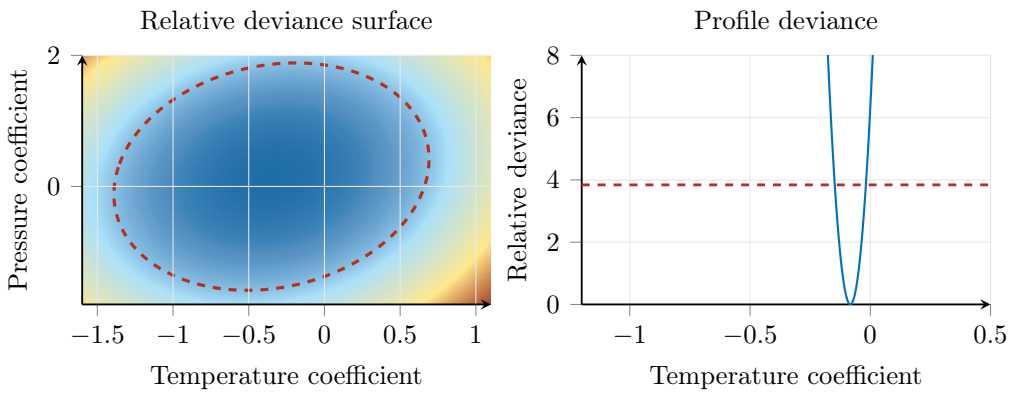

Figure 6

occurs, then

$$\mathbb{P}(Z_{ij} = 1) = 1 - \mathbb{P}(K_{ij} = 0) = 1 - \exp(-\delta\lambda_i).$$

Writing this probability as μ_i yields

$$\log(-\log(1 - \mu_i)) = \log(\delta) + \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

This is the **complementary log-log link**. Its coefficients are log rate ratios rather than log odds ratios, and $\log(\delta)$ is an offset. The logit link remains the standard choice unless a mechanism such as an underlying event process motivates a different link; finite data often cannot discriminate reliably among plausible links.

Continuous Responses and Model Diagnostics

Positive continuous responses may also be modelled within or alongside the generalized linear model framework. The cricketer data provide an example. The response is lifetime, and the explanatory variables include birth year and handedness. The motivating question is whether left-handed people died younger, perhaps because of accidents involving equipment designed for right-handed people. This is fundamentally a survival problem, but several continuous response models clarify the choices involved.

Definition 2.9 A positive variable Y has a **gamma distribution** with scale $\phi > 0$ and shape $\nu > 0$ if

$$f(y; \phi, \nu) = \frac{(y/\phi)^{\nu-1} \exp(-y/\phi)}{\Gamma(\nu)\phi}, \quad y > 0.$$

Its mean and variance are

$$\mathbb{E}[Y] = \phi\nu \qquad \text{and} \qquad \operatorname{Var}(Y) = \phi^2\nu.$$

A gamma regression parameterized by its mean uses

$$Y_i \sim \text{Gamma}(\mu_i/\nu, \nu) \qquad \text{and} \qquad \log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

In this parameterization, the scale is μ_i/ν and the dispersion is $1/\nu$. For cricketers born before 1890 and not killed as soldiers, the fitted coefficients are

$$\widehat{\beta}_0 = 4.1751, \qquad \widehat{\beta}_{\text{decade}} = 0.0242, \qquad \text{and} \qquad \widehat{\beta}_{\text{left}} = -0.0162,$$

with estimated shape 18.7245. The handedness coefficient has standard error 0.0136 and $p = 0.2341$. A histogram for right-handed cricketers born near 1850 shows that the fitted gamma distribution does not capture the empirical right skew well.

Three alternatives are useful. A Weibull regression is standard for event times and has a direct survival interpretation. A **Gompertz model** permits a hazard that changes exponentially with age. A **Box–Cox transformation** instead transforms the response before applying a Gaussian linear model. For these data, the profile likelihood favours a transformation close to Y^2.

The fitted gamma, Weibull, Gompertz, and squared response Gaussian densities are compared with the lifetime histogram in Figure 7. The figure emphasizes why distributional choice should be assessed on the response scale rather than selected only from coefficient tables.

In the Weibull fit, the estimated percentage change in expected lifetime for left-handed rather than right-handed cricketers is -0.72%, with an approximate 95% interval from -2.88% to 1.50%. The data do not provide evidence of shorter expected lifetime for left-handed cricketers after allowing for birth decade.

Continuous response families such as gamma and Weibull are not nested, so an ordinary likelihood ratio test cannot select among them. Generalized linear models for continuous data are consequently less routine than binomial and Poisson models. Transforming the response toward Gaussianity may produce a simpler model, while a Weibull model often gives parameters that are easier to interpret for event times.

Interactions require careful contrasts on the natural scale. Suppose that a log link model contains birth decade, handedness, cause of death, and an interaction between handedness and cause of

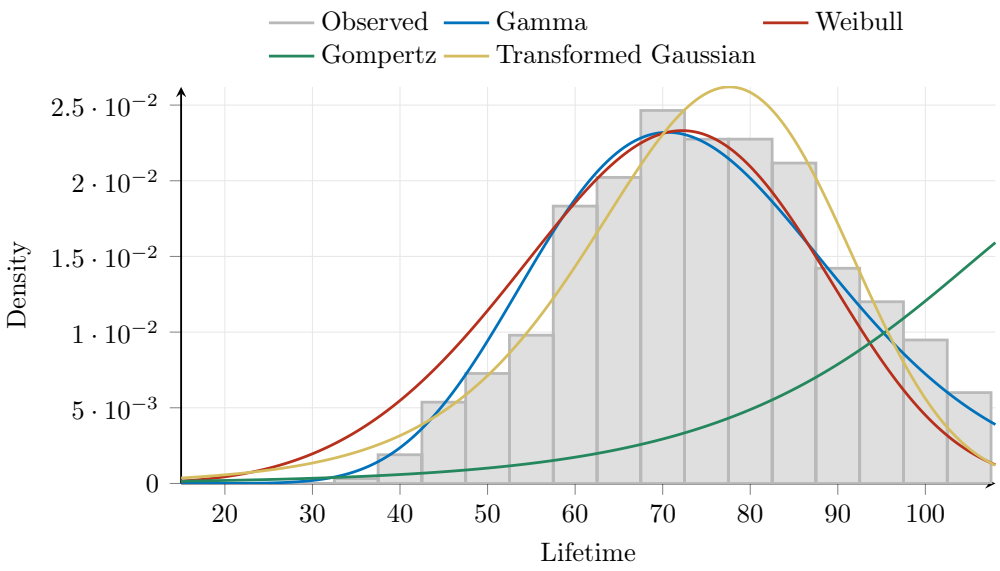

Figure 7

death:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i,\text{decade}} + \beta_2 x_{i,\text{left}} + \beta_3 x_{i,\text{inbed}} + \beta_4 x_{i,\text{left}} x_{i,\text{inbed}}.$$

Then $\exp(\beta_2)$ compares left- and right-handed cricketers among accidental deaths, $\exp(\beta_3)$ compares deaths in bed with accidental deaths among right-handed cricketers, and

$$\exp(\beta_3 + \beta_4)$$

is the ratio of expected lifetimes for deaths in bed versus accidental deaths among left-handed cricketers. More generally, if the desired contrast is $\boldsymbol{a}^\top \boldsymbol{\beta}$, then

$$\boldsymbol{a}^\top \widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\left(\boldsymbol{a}^\top \boldsymbol{\beta}, \boldsymbol{a}^\top \mathbf{I}(\widehat{\boldsymbol{\beta}})^{-1} \boldsymbol{a}\right).$$

This formula supplies standard errors for scientifically meaningful contrasts that are not individual fitted coefficients.

Diagnostics are more difficult for binary and count responses because residuals are discrete and heteroscedastic. Exploratory plots, comparisons of empirical and fitted distributions, binned residual displays, and checks for overdispersion or zero inflation can still reveal important failures. A model should respect the support of the response, reproduce its principal distributional features, and answer the scientific question on a scale that can be interpreted.

Lecture 3 — Wednesday, January 23, 2019

3. Mixed Models and Bayesian Computation

Linear Mixed Models

Repeated measurements on one individual, children attending one school, and specimens taken from one animal are not independent. A **mixed model** represents the common information shared by observations from the same unit through latent random effects.

Definition 3.1 In the Gaussian **random intercept model**,

$$Y_{ij} \mid U_i \sim \mathcal{N}\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + U_i, \tau^2\right) \qquad \text{and} \qquad U_i \sim \mathcal{N}(0, \sigma^2).$$

Conditional on the random effects, the responses are independent, and the U_i are independent and identically distributed. The term $\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$ contains the **fixed effects**, U_i is the **random effect** for unit i, and τ^2 is the observation level variance.

The random intercept is the unit's deviation from the population mean after accounting for the fixed effects. Equivalently,

$$Y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \varepsilon_{ij} \qquad \text{and} \qquad \varepsilon_{ij} = U_i + Z_{ij},$$

where the Z_{ij} are independent and identically distributed as $\mathcal{N}(0, \tau^2)$. The marginal errors are Gaussian but correlated:

$$\mathrm{Var}(\varepsilon_{ij}) = \sigma^2 + \tau^2 \qquad \text{and} \qquad \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = \sigma^2 \quad \text{for } j \neq k.$$

The resulting correlation within a unit is

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{\sigma^2}{\sigma^2 + \tau^2}.$$

Observations from different units remain independent.

The right panel of Figure 8 shows random intercept dependence: a subject who is above the population trend at one age tends to remain above it. The left panel shows observations generated

independently around the common mean.

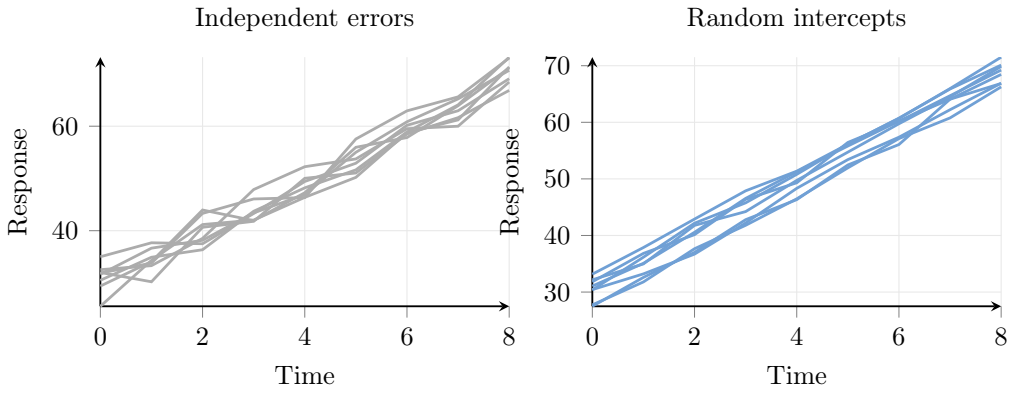

Figure 8

Random intercepts describe only one kind of dependence. A **random slope** allows the separation between subjects to grow or contract with a covariate, while **serial correlation** describes errors that evolve gradually through time. These structures may look similar in short series, so scientific knowledge and the implied covariance matrix matter.

Example 3.2 The pigs data contain weights for 48 pigs at 9 time points. A random intercept model is

$$Y_{ij} \mid U_i \sim \mathcal{N}(\beta_0 + \beta_1 t_{ij} + U_i, \tau^2) \qquad \text{and} \qquad U_i \sim \mathcal{N}(0, \sigma^2).$$

Here, β_0 is the population mean weight at time zero, β_1 is the population mean gain per time unit, and U_i is pig i's persistent deviation from that population trajectory.

The individual trajectories in Figure 9 are approximately parallel, which makes a random intercept a plausible first model.

The fit is specified by the concise R command

```
fit <- nlme::lme(weight ~ time, random = ~1 | id, data = pigs)
```

and gives

$$\widehat{\beta}_0 = 19.36, \qquad \widehat{\beta}_1 = 6.21, \qquad \widehat{\sigma} = 3.89, \qquad \text{and} \qquad \widehat{\tau} = 2.10.$$

The estimated standard errors of the fixed intercept and slope are 0.60 and 0.04, respectively.

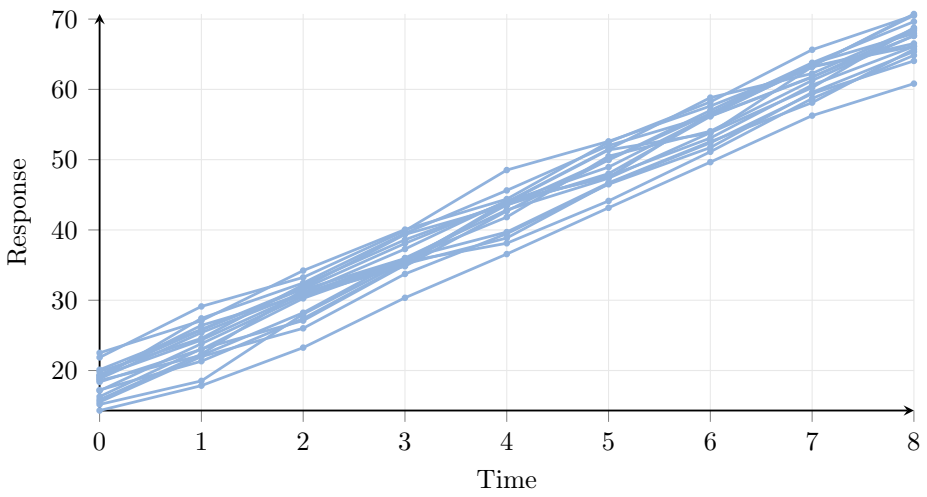

Figure 9

The terms **random effects model**, **latent variable model**, **multilevel model**, and **hierarchical model** emphasize different aspects of the same construction. A model is not Bayesian merely because it contains latent random effects; the inferential framework depends on whether the unknown parameters are assigned **prior distributions**.

Likelihood, Restricted Likelihood, and Prediction

Let $\boldsymbol{y}$ stack all $N = \sum_{i=1}^{M} J_i$ observations, let $\mathbf{X}$ be the $N \times P$ fixed effect design matrix with full column rank, and let $\mathbf{A}$ be the $M \times N$ incidence matrix with $A_{ik} = 1$ when observation k belongs to unit i. In the general Gaussian random intercept model,

$$\boldsymbol{y} \mid \boldsymbol{U} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{A}^{\top}\boldsymbol{U}, \tau^2 \mathbf{I}_N\right) \qquad \text{and} \qquad \boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

For independent random intercepts, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_M$.

Integrating out $\boldsymbol{U}$ gives the **marginal model**

$$\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}\right) \qquad \text{and} \qquad \mathbf{V} = \mathbf{A}^{\top}\boldsymbol{\Sigma}\mathbf{A} + \tau^2 \mathbf{I}_N. \tag{3.1}$$

Thus, random effects are one structured way to construct correlated marginal errors.

For a multivariate Gaussian vector, the log likelihood from (3.1) satisfies

$$-2\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \tau; \boldsymbol{y}) = \log|\mathbf{V}| + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + C.$$

Direct inversion is neither necessary nor numerically desirable. Three computational ideas make the likelihood practical.

First, $\mathbf{V}$ is often sparse or has sparse components. A **Cholesky factorization** $\mathbf{V} = \mathbf{L}\mathbf{L}^\top$ reduces quadratic forms to triangular solves and computes the determinant from the diagonal entries of $\mathbf{L}$.

Second, write

$$\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}/\tau^2 \qquad \text{and} \qquad \widetilde{\mathbf{V}} = \mathbf{A}^\top \widetilde{\boldsymbol{\Sigma}} \mathbf{A} + \mathbf{I}_N.$$

For a fixed variance ratio $\widetilde{\boldsymbol{\Sigma}}$, **generalized least squares** gives

$$\widehat{\boldsymbol{\beta}}(\widetilde{\boldsymbol{\Sigma}}) = \left(\mathbf{X}^\top \widetilde{\mathbf{V}}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \widetilde{\mathbf{V}}^{-1} \boldsymbol{y}$$

and

$$\widehat{\tau}^2(\widetilde{\boldsymbol{\Sigma}}) = \frac{\left\{\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\}^\top \widetilde{\mathbf{V}}^{-1} \left\{\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\}}{N}.$$

These expressions profile $\boldsymbol{\beta}$ and τ^2 out of the numerical optimization.

Third, when $\widetilde{\boldsymbol{\Sigma}}$ is positive definite, the **Woodbury identity** and **matrix determinant lemma** reduce an $N \times N$ calculation to an $M \times M$ calculation when $M \ll N$:

$$\left(\mathbf{I}_N + \mathbf{A}^\top \widetilde{\boldsymbol{\Sigma}} \mathbf{A}\right)^{-1} = \mathbf{I}_N - \mathbf{A}^\top \left(\widetilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{A}\mathbf{A}^\top\right)^{-1} \mathbf{A}$$

and

$$\left|\mathbf{I}_N + \mathbf{A}^\top \widetilde{\boldsymbol{\Sigma}} \mathbf{A}\right| = \left|\widetilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{A}\mathbf{A}^\top\right| \left|\widetilde{\boldsymbol{\Sigma}}\right|.$$

Ordinary maximum likelihood tends to underestimate variance components because the residuals use an estimated $\boldsymbol{\beta}$. **Restricted maximum likelihood** removes the fixed effect directions before estimating the covariance parameters. Let $\mathbf{K}$ be an $N \times (N - P)$ matrix of full column rank satisfying $\mathbf{K}^\top \mathbf{X} = \mathbf{0}$, and define

$$\boldsymbol{y}^* = \mathbf{K}^\top \boldsymbol{y}.$$

Then

$$\boldsymbol{y}^* \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}^\top \mathbf{V} \mathbf{K}\right),$$

so its likelihood contains no $\boldsymbol{\beta}$. The identities of Searle and Quaas yield the restricted log likelihood

$$-2\ell_R(\mathbf{V}; \boldsymbol{y}) = \log|\mathbf{V}| + \log\left|\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right| + \left\{\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{V})\right\}^\top \mathbf{V}^{-1} \left\{\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\mathbf{V})\right\} + C. \tag{3.2}$$

This expression does not depend on the particular choice of $\mathbf{K}$.

After profiling over τ^2, restricted maximum likelihood uses

$$\widehat{\tau}^2_R(\widetilde{\boldsymbol{\Sigma}}) = \frac{\left\{\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\}^\top \widetilde{\mathbf{V}}^{-1} \left\{\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\}}{N - P}.$$

When $\mathbf{V} = \tau^2 \mathbf{I}_N$, this reduces to the familiar residual sum of squares divided by $N-P$. Maximum likelihood and restricted maximum likelihood are often similar when P is small and the covariance matrix is well conditioned. When they differ materially, the restricted estimates of variance components are generally preferable.

If $\mathbf{V}$ were known, then

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \boldsymbol{y}$$

would be unbiased with variance

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^{-1}. \tag{3.3}$$

In practice, estimated covariance parameters are substituted into (3.3). This **plug-in approximation** ignores part of the uncertainty in estimating $\boldsymbol{\Sigma}$ and τ^2.

The random effects are predicted from their conditional Gaussian distribution. Since

$$\begin{pmatrix} \boldsymbol{U} \\ \boldsymbol{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{X}\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\mathbf{A} \\ \mathbf{A}^\top \boldsymbol{\Sigma} & \mathbf{V} \end{pmatrix} \right),$$

the conditional distribution is

$$\boldsymbol{U} \mid \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\Sigma}\mathbf{A}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}\mathbf{V}^{-1}\mathbf{A}^\top\boldsymbol{\Sigma} \right). \tag{3.4}$$

The conditional mean in (3.4) is a compromise between the data for that unit and the population mean. If a unit has no observations, its conditional distribution equals its marginal distribution. Frequentist software ordinarily substitutes $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Sigma}}$, and $\widehat{\tau}^2$ into this distribution and does not fully propagate their estimation uncertainty.

For a future observation Y_{ik} on an already observed unit,

$$\mathbb{E}[Y_{ik} \mid \boldsymbol{y}] = \boldsymbol{x}_{ik}^\top \boldsymbol{\beta} + \mathbb{E}[U_i \mid \boldsymbol{y}]$$

and

$$\mathrm{Var}(Y_{ik} \mid \boldsymbol{y}) = \mathrm{Var}(U_i \mid \boldsymbol{y}) + \tau^2.$$

Prediction for a new unit uses the population distribution of its random effect rather than a conditional distribution specific to that unit.

Longitudinal and Multilevel Models

The London school data illustrate how random effects separate levels of variation. The response is an examination score, and the fixed effects include sex and ethnicity. Raw means for ethnic groups range from about 17.1 for Caribbean students to 27.1 for Southeast Asian students, but these differences may partly reflect the schools attended. A random school intercept allows the model to compare groups while accounting for persistent differences among schools.

The fitted school and individual standard deviations are

$$\widehat{\sigma} = 3.93 \qquad \text{and} \qquad \widehat{\tau} = 11.84.$$

The fixed effect estimates relative to a white male student are shown below.

effect	estimate	standard error	effect	estimate	standard error
intercept	18.88	0.39	female	3.00	0.27
African	2.38	0.64	Arab	1.00	1.43
Bangladeshi	−0.35	0.82	Caribbean	−2.83	0.27
Greek	2.50	0.86	Indian	5.49	0.62
Pakistani	4.62	0.80	Southeast Asian	6.93	0.81
Turkish	−0.97	0.75	other	2.43	0.43

Individual variation is much larger than the fitted school variation, but school effects remain comparable with several contrasts among ethnic groups. Two students with identical fixed effects in the same school have a score difference with standard deviation $\sqrt{2}\widehat{\tau} \approx 16.74$. The coefficients describe adjusted associations and should not be interpreted as intrinsic group effects.

A school may instead be represented by a separate fixed parameter θ_i. Fixed school effects make no distributional assumption and can be sensible when only a few schools are of interest. Random effects use fewer parameters, permit inference about a broader population of schools, and predict school effects through (3.4). Their conditional means exhibit **shrinkage**: imprecisely estimated school effects are pulled toward the population mean. Shrinkage is stronger when a school has fewer students or when the variance among schools is small.

Figure 10 compares separate fixed effect estimates with conditional random effect predictions. The latter are closer to the overall mean because they combine the school data with the fitted population distribution.

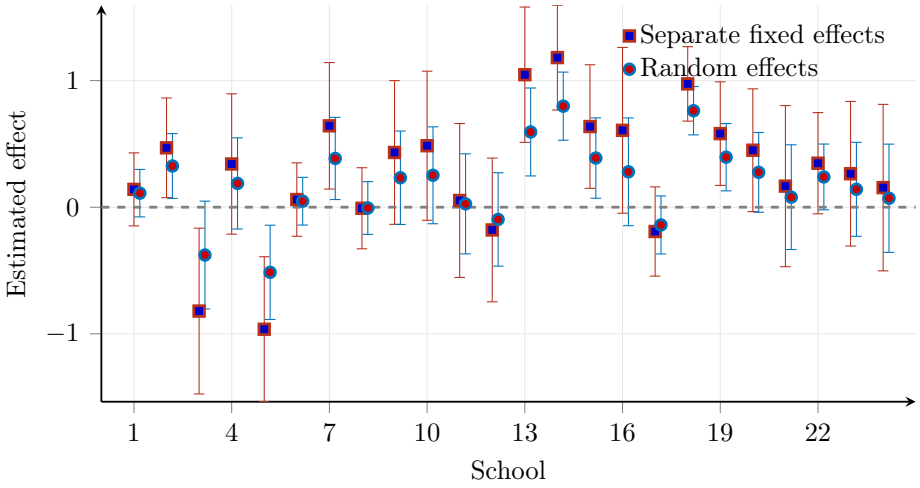

Figure 10

Estimating a variance from only three or four groups is unreliable, so fixed effects are often safer with very few groups. With a moderate or large number of exchangeable groups, a random effect model is usually more economical and more useful for population inference.

The milk protein data contain repeated measurements from 79 cows assigned to barley, lupin, or mixed diets. The observed trajectories are shown in Figure 11.

Treating time as categorical assigns a separate mean to each observed time without imposing a parametric trend. The additive random intercept model

$$Y_{ij} = \theta_j + \alpha_{\text{lupins}} L_i + \alpha_{\text{mixed}} M_i + U_i + Z_{ij}$$

assumes that all three diets share the same time profile, shifted vertically by constant diet effects. Its fitted baseline is approximately 3.94; the lupin and mixed effects are -0.20 and -0.10. The fitted random intercept and residual standard deviations are 0.17 and 0.25.

The fully interacted model assigns a separate mean to every combination of diet and time. It is flexible but uses 59 parameters rather than 23. A maximum likelihood comparison with the additive model gives a likelihood ratio statistic of 24.794 on 36 degrees of freedom, with $p = 0.9205$. There is no evidence that a completely unrestricted diet contrast is needed at every time point.

A useful intermediate model leaves the baseline barley trajectory categorical but describes each

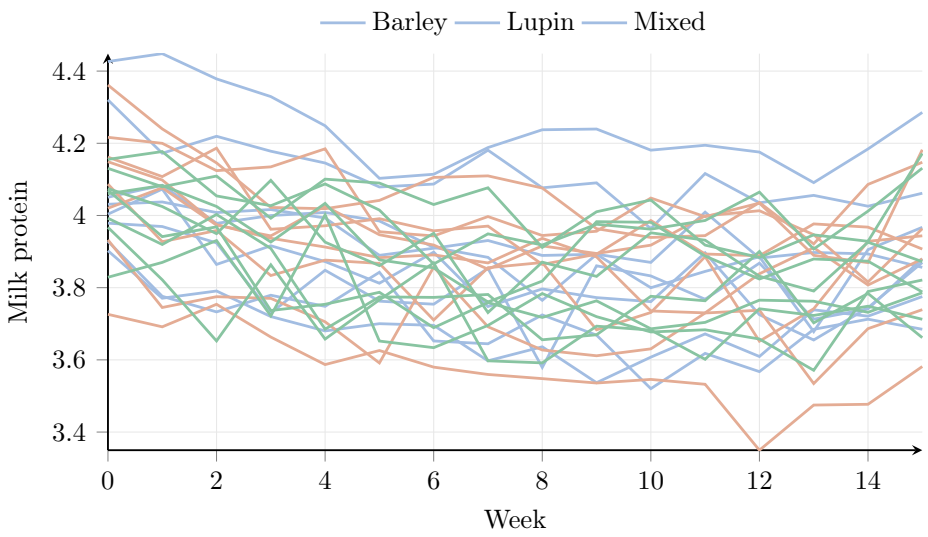

Figure 11

treatment contrast by a quadratic function of time:

$$\mathbb{E}[Y_{ij}] = \begin{cases} \theta_j, & \text{if cow } i \text{ receives barley,} \\ \theta_j + \alpha_1 + \beta_1 t_j + \gamma_1 t_j^2, & \text{if cow } i \text{ receives lupins,} \\ \theta_j + \alpha_2 + \beta_2 t_j + \gamma_2 t_j^2, & \text{if cow } i \text{ receives the mixed diet.} \end{cases}$$

Adding the two linear contrast terms to the additive model gives a likelihood ratio statistic of 7.4573 on two degrees of freedom, with $p = 0.0240$. Adding the two quadratic terms gives only 0.0905, with $p = 0.9558$. The linear contrast model therefore captures the supported treatment difference over time without the full interaction.

Likelihood ratio tests comparing fixed effect structures must use maximum likelihood rather than restricted maximum likelihood because the restricted likelihood changes when $\mathbf{X}$ changes. Testing whether a random effect variance is zero is different: zero lies on the boundary of the parameter space. Under the usual regularity conditions for a single scalar variance component, the asymptotic null distribution is a mixture with probability one half at zero and probability one half on a χ^2_1 distribution.

A random slope model permits responses specific to each unit:

$$Y_{ij} \mid \boldsymbol{U}_i \sim \mathcal{N}\left(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + U_{i1} + U_{i2} w_{ij}, \tau^2\right) \qquad \text{and} \qquad \begin{pmatrix} U_{i1} \\ U_{i2} \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}).$$

The random intercept and slope may be correlated. Stacking the pair for every unit gives a **block**

diagonal covariance matrix

$$\boldsymbol{\Sigma} = \mathbf{I}_M \otimes \boldsymbol{\Gamma},$$

where $\otimes$ is the **Kronecker product**. The marginal model retains the form in (3.1) with an expanded random effect design matrix.

For the cow data, a random intercept and time slope give residual standard deviation 0.22199. Relative to that residual standard deviation, the random intercept and random slope standard deviations are 1.12585 and 0.11114, with correlation -0.734. Cows with higher initial protein measurements therefore tend to have more negative time slopes under this fitted model.

The additive and contrast fits are compared in Figure 12. The common time pattern is nonlinear, while the supported differences between diets are much simpler.

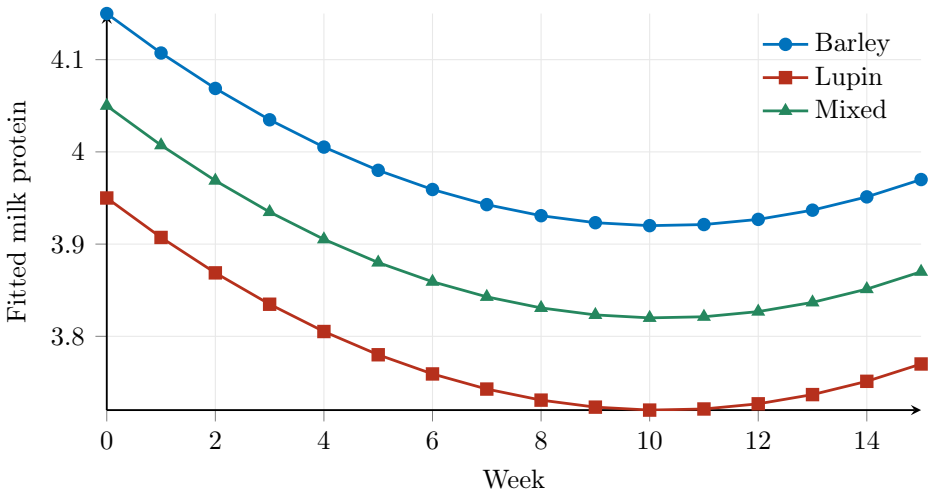

Figure 12

Several nested random effects can be included. If measurement ℓ belongs to student k, class j, and school i, then a model with four levels is

$$Y_{ijk\ell} = \boldsymbol{x}_{ijk\ell}^\top \boldsymbol{\beta} + B_i + C_{ij} + D_{ijk} + Z_{ijk\ell},$$

where

$$B_i \sim \mathcal{N}(0, \sigma_B^2), \qquad C_{ij} \sim \mathcal{N}(0, \sigma_C^2), \qquad D_{ijk} \sim \mathcal{N}(0, \sigma_D^2), \qquad \text{and} \qquad Z_{ijk\ell} \sim \mathcal{N}(0, \tau^2),$$

independently. The covariance matrix is a sum of contributions from the incidence matrix at each level:

$$\mathbf{V} = \mathbf{A}_B^\top \boldsymbol{\Sigma}_B \mathbf{A}_B + \mathbf{A}_C^\top \boldsymbol{\Sigma}_C \mathbf{A}_C + \mathbf{A}_D^\top \boldsymbol{\Sigma}_D \mathbf{A}_D + \tau^2 \mathbf{I}_N.$$

For a school leavers model with tests nested within students, students within classes, and classes within schools, the fitted residual standard deviation is 4.663. The fitted standard deviations are approximately 0.003 between schools, 2.173 between classes within schools, and 5.527 between students within classes. The school component is essentially zero after accounting for the other levels and fixed effects. Social class estimates relative to class I range from approximately -0.11 for class II to -5.17 for class V, while the estimated male effect is -0.13.

Lecture 4 — Wednesday, January 30, 2019

Bayesian Hierarchical Models

Bayesian inference treats the unknown parameter as a random quantity and updates a prior distribution with the observed likelihood. Consider $Y = 4$ diseased children in a sample of $N = 12$.

Example 4.1 Suppose that

$$Y \mid \theta \sim \text{Bin}(12, \theta) \qquad \text{and} \qquad \theta \sim \text{Beta}(1.5, 2).$$

Bayes' rule gives the **posterior distribution**:

$$\pi(\theta \mid Y = 4) = \frac{\mathbb{P}(Y = 4 \mid \theta)\pi(\theta)}{\int_0^1 \mathbb{P}(Y = 4 \mid u)\pi(u)\,\mathrm{d}u} \propto \mathbb{P}(Y = 4 \mid \theta)\pi(\theta),$$

and **beta-binomial conjugacy** yields

$$\theta \mid Y = 4 \sim \text{Beta}(5.5, 10).$$

The posterior probability of prevalence at least one half is

$$\mathbb{P}(\theta \geq 0.5 \mid Y = 4) = 0.1182.$$

By comparison, the one sided frequentist tail probability under $\theta = 0.5$ is

$$\mathbb{P}(Y \leq 4 \mid \theta = 0.5) = 0.1938.$$

These probabilities answer different questions.

The prior and posterior from Example 4.1 are shown in Figure 13. The posterior is more concen-

trated because it combines the prior information with the sample.

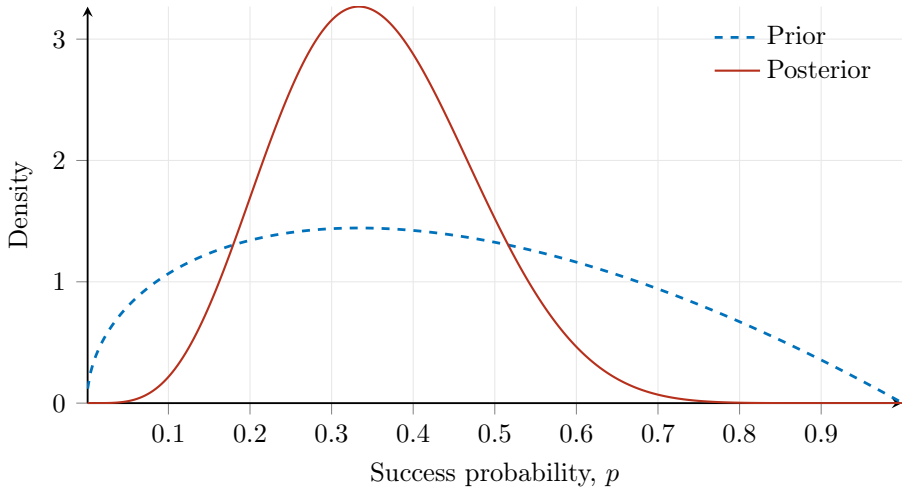

Figure 13

In general, a prior $\theta \sim \text{Beta}(\nu_1, \nu_2)$ and an observation $Y = y$ from $\text{Bin}(N, \theta)$ give

$$\theta \mid Y = y \sim \text{Beta}(\nu_1 + y, \nu_2 + N - y).$$

Different priors produce different posterior probabilities. With the same data, the priors $\text{Beta}(1, 5)$, $\text{Beta}(1, 1)$, and $\text{Beta}(5, 1)$ give posterior probabilities

$$0.0245, \qquad 0.1334, \qquad \text{and} \qquad 0.5000$$

that $\theta \geq 0.5$. This dependence is not a defect to be hidden: it makes the assumptions explicit and invites **sensitivity analysis**. As the sample size increases, the likelihood usually dominates reasonable fixed priors. If one third of sampled children are diseased and N is 3, 12, or 24, the corresponding probabilities under the prior in Example 4.1 are 0.2639, 0.1182, and 0.0484.

Definition 4.2 For a parameter with a continuous posterior distribution, an interval (a, b) is a $100(1 - \alpha)\%$ **posterior credible interval** for θ if

$$\mathbb{P}(a < \theta < b \mid \boldsymbol{Y}) = 1 - \alpha.$$

An **equal tailed interval** additionally satisfies

$$\mathbb{P}(\theta < a \mid \boldsymbol{Y}) = \mathbb{P}(\theta > b \mid \boldsymbol{Y}) = \alpha/2.$$

The equal tailed 95% and 99% credible intervals in Example 4.1 are

$$(0.1457, 0.5994) \qquad \text{and} \qquad (0.1018, 0.6738),$$

respectively. A credible interval assigns posterior probability to the parameter after observing the data. A frequentist confidence procedure instead has repeated sampling coverage: the random interval covers a fixed parameter in a specified proportion of hypothetical repetitions. Once a particular confidence interval has been observed, the fixed parameter is either inside it or outside it.

Prior information should be formulated on a scientifically meaningful scale. A uniform prior for a probability is not uniform for its odds or log odds. A prior that appears weak for θ may be strong after transforming to $\text{logit}(\theta)$. **Informative priors** should come from defensible external knowledge, ideally elicited before examining the current outcomes. **Weakly informative priors** constrain implausible regions without claiming precise knowledge.

Hierarchical models make this updating principle especially useful. The bacteria data record whether a blood sample contains a particular bacterium at several weeks for each individual. Let $Y_{it} = 1$ denote presence for individual i at time t. A logistic random intercept model is

$$Y_{it} \mid U_i \sim \text{Bernoulli}(\lambda_{it}), \qquad \text{logit}(\lambda_{it}) = \mu + \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + U_i, \qquad \text{and} \qquad U_i \sim \mathcal{N}(0, \sigma^2). \tag{4.1}$$

Conditional on the random effects, the responses are independent. The explanatory variables represent week and active treatment. A positive U_i identifies an individual who is more susceptible than average and induces dependence among that individual's observations.

A representative prior specification is

$$\mu \sim \mathcal{N}(0, 10^2), \qquad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 3^2\mathbf{I}), \qquad \text{and} \qquad \sigma \sim \text{Exp}(4\log(2)).$$

The last choice has prior median $1/4$. The quantities μ, $\boldsymbol{\beta}$, and σ are parameters; the U_i are **latent variables**. Numerical constants controlling the priors are **hyperparameters**. The presence of random effects does not itself determine whether inference is Bayesian or frequentist.

Frequentist likelihood inference integrates out the random effects:

$$L(\mu, \boldsymbol{\beta}, \sigma; \boldsymbol{y}) = \int_{\mathbb{R}^M} \mathbb{P}(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{U}; \mu, \boldsymbol{\beta}) \pi(\boldsymbol{U}; \sigma) \, \mathrm{d}\boldsymbol{U}.$$

For non-Gaussian responses, this integral is generally intractable. Penalized quasi-likelihood, generalized estimating equations, and hierarchical likelihood methods can be fast, but may perform poorly with strongly non-Gaussian responses or few observations per cluster. Laplace approxima-

tion and automatic differentiation provide better likelihood based alternatives in many problems.

Bayesian inference instead works with the joint posterior

$$\pi(\boldsymbol{U}, \mu, \boldsymbol{\beta}, \sigma \mid \boldsymbol{y}) \propto \mathbb{P}(\boldsymbol{y} \mid \boldsymbol{U}, \mu, \boldsymbol{\beta}) \pi(\boldsymbol{U} \mid \sigma) \pi(\sigma) \pi(\mu) \pi(\boldsymbol{\beta}).$$

Marginal posterior distributions are obtained by integrating out every other component. For example,

$$\pi(U_1 \mid \boldsymbol{y}) = \int \pi(\boldsymbol{U}, \mu, \boldsymbol{\beta}, \sigma \mid \boldsymbol{y}) \, \mathrm{d}\mu \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\sigma \, \mathrm{d}U_2 \cdots \mathrm{d}U_M.$$

This integration propagates uncertainty in the regression and variance parameters into inference for each random effect.

For the priors above, the bacteria analysis gives a posterior treatment mean of -1.15, posterior standard deviation 0.50, and 95% credible interval $(-2.19, -0.21)$. The posterior probability that the treatment coefficient is positive is 0.0059, so the treatment is very likely to reduce bacterial presence on the log odds scale. The posterior means of the individual random effects vary, although their individual 95% credible intervals all include zero.

The left panel of Figure 14 compares the prior and posterior of σ; the right panel shows the marginal posterior of the treatment coefficient. The variance component remains sensitive to the prior because the likelihood contains limited information near $\sigma = 0$.

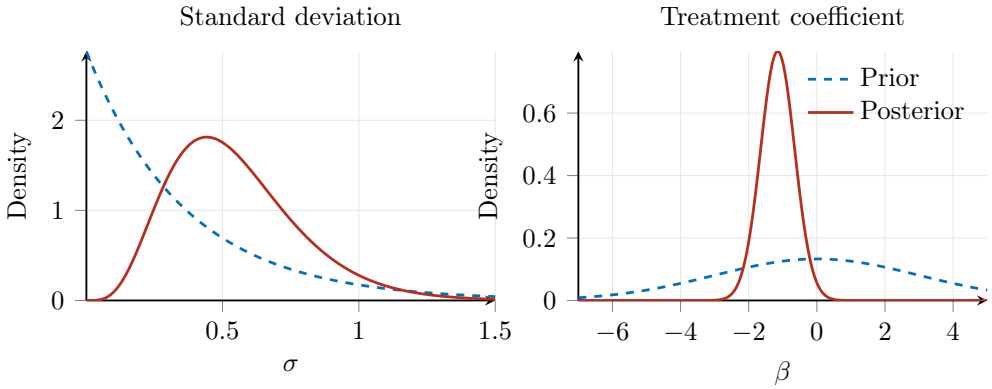

Figure 14

Laplace Approximation and Integrated Nested Laplace Approximation

Conjugate analysis is available only for special pairs of prior and likelihood distributions. Markov chain Monte Carlo methods simulate from complicated posteriors without evaluating the normalizing constant, but they may be computationally expensive and require careful convergence

assessment. Integrated nested Laplace approximation provides fast deterministic marginal inference for latent Gaussian models. The foundational method is described by Rue, Martino, and Chopin.

Definition 4.3 A **latent Gaussian model** has conditionally independent responses

$$Y_i \mid \eta_i, \boldsymbol{\theta} \sim f(y_i \mid \eta_i, \boldsymbol{\theta}) \qquad \text{and} \qquad g(\lambda_i) = \eta_i,$$

with a linear latent predictor

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{a}_i^\top \boldsymbol{U}.$$

Conditional on a low dimensional hyperparameter vector $\boldsymbol{\theta}$, the latent variables and regression coefficients are jointly Gaussian:

$$\boldsymbol{U} \mid \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta})) \qquad \text{and} \qquad \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}).$$

The **latent field**

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\beta} \\ \boldsymbol{U} \end{pmatrix}$$

is Gaussian given $\boldsymbol{\theta}$, often with a sparse **precision matrix** $\mathbf{Q}(\boldsymbol{\theta})$. The dimension of $\boldsymbol{W}$ may be very large, while $\boldsymbol{\theta}$ must remain modest. Sparsity is the principal computational resource.

Definition 4.4 Let a density be proportional to $\exp(\ell(\boldsymbol{x}))$, with a unique mode $\boldsymbol{x}^*$. If

$$\mathbf{H} = -\nabla^2 \ell(\boldsymbol{x})\big|_{\boldsymbol{x}=\boldsymbol{x}^*}$$

is positive definite, the **Laplace approximation** replaces the density locally by

$$\boldsymbol{X} \dot{\sim} \mathcal{N}\left(\boldsymbol{x}^*, \mathbf{H}^{-1}\right).$$

The approximation follows from a second order Taylor expansion of the log density. It is accurate when the density is sufficiently smooth, unimodal, and close to Gaussian near its dominant mass.

Integrated nested Laplace approximation (INLA) proceeds in three nested stages.

1. For fixed $\boldsymbol{\theta}$, find the mode $\boldsymbol{W}^*(\boldsymbol{\theta})$ of $\pi(\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\theta})$ and construct a Gaussian or Laplace approximation $\widetilde{\pi}_G(\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\theta})$.

2. Approximate the hyperparameter posterior by

$$\widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{W} \mid \boldsymbol{\theta})\mathbb{P}(\boldsymbol{y} \mid \boldsymbol{W}, \boldsymbol{\theta})}{\widetilde{\pi}_G(\boldsymbol{W} \mid \boldsymbol{y}, \boldsymbol{\theta})} \right|_{\boldsymbol{W}=\boldsymbol{W}^*(\boldsymbol{\theta})}.$$

3. Numerically integrate over $\boldsymbol{\theta}$ to obtain marginal posteriors such as

$$\widetilde{\pi}(U_j \mid \boldsymbol{y}) = \int \widetilde{\pi}(U_j \mid \boldsymbol{y}, \boldsymbol{\theta})\widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}.$$

The integration in (3) and the nested approximation in (1) explain the method's name. No Gaussian approximation is imposed directly on $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$. A more accurate Laplace or simplified Laplace approximation can replace the Gaussian approximation for individual latent marginals.

For the bacteria model in (4.1), the essential syntax is

```
fit <- INLA::inla(
  newy ~ factor(week) + treatment + f(ID, model = "iid"),
  family = "binomial", data = bacteria
)
```

The fitted object contains summaries and numerical marginal densities for fixed effects, random effects, fitted values, and hyperparameters. Posterior quantiles transform directly under a monotone function, with their order reversed for a decreasing transformation. Posterior means do not generally transform by applying the function to the original posterior mean.

Variance components are usually reported as standard deviations rather than precisions. A prior on the precision $1/\sigma^2$ can behave very differently from an apparently similar prior on σ. Default gamma priors on precision may place little mass near the simpler model $\sigma = 0$, and the bacteria example changes drastically under different such defaults.

A **penalized complexity (PC) prior** starts from a simpler base model and penalizes departure from it. For a Gaussian random effect, the base model is $\sigma = 0$. The construction of Simpson and coauthors places an exponential prior on the distance from that base. It can be calibrated by

$$\mathbb{P}(\sigma > u) = \alpha,$$

which gives exponential rate $-\log(\alpha)/u$. This statement is usually easier to elicit and interpret than gamma shape and rate parameters for the precision.

The three variance component priors in Figure 15 have visibly different behaviour near the base model. The exponential prior on σ can shrink toward zero, whereas the gamma priors on precision may resist that limit.

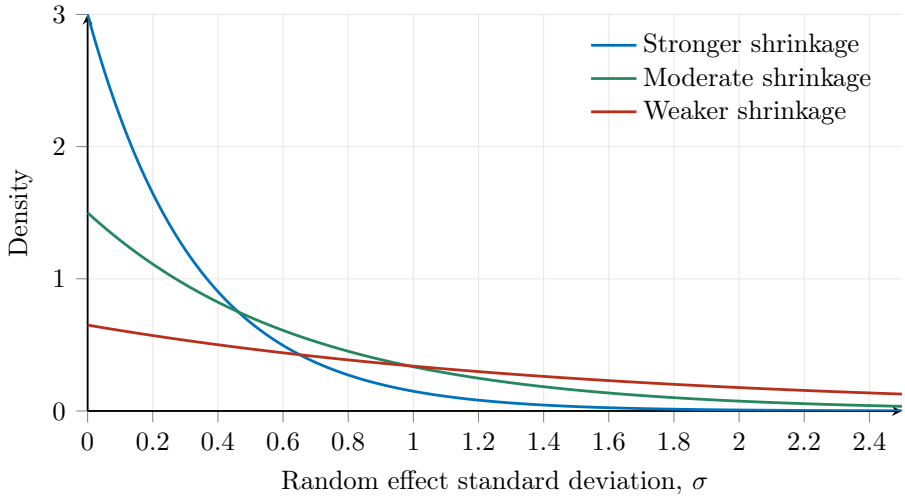

Figure 15

Example 4.5 The Gibraltar data contain 44,275 daily death counts over approximately 140 years. The question is whether the levanter wind is associated with mortality after accounting for population, long term trend, day of the week, seasonal cycles, overdispersion, and temporal dependence. Let P_i denote population exposure. A latent Gaussian model is

$$Y_i \mid U(t_i), Z_i \sim \text{Poisson}(\lambda_i P_i),$$

$$\log(\lambda_i) = \mu + \boldsymbol{x}_i^\top \boldsymbol{\beta} + U(t_i) + Z_i,$$

where

$$\text{Cov}(U(t), U(t+h)) = \sigma^2 \exp(-|h|/\phi) \qquad \text{and} \qquad Z_i \sim \mathcal{N}(0, \tau^2).$$

The smooth process $U(t)$ describes temporally correlated logarithmic relative risk, while Z_i represents independent extra-Poisson variation.

On an equally spaced daily grid, the exponential covariance can be represented by a first order autoregressive process

$$U_1 \sim \mathcal{N}\left(0, \frac{1}{\kappa(1-\rho^2)}\right), \qquad U_i = \rho U_{i-1} + \varepsilon_i, \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \kappa^{-1}),$$

with $|\rho| < 1$. For one day spacing,

$$\rho = \exp(-1/\phi).$$

A prior favouring a smooth process with low variance places mass near $\sigma = 0$ and $\rho = 1$, while allowing the data to support rougher temporal variation.

The fitted levanter coefficient is -0.0003, with 95% credible interval $(-0.0042, 0.0036)$, so there is no evidence of an association with mortality. The log rate trend per decade is -0.0766, with interval $(-0.0824, -0.0708)$. Seasonal sine and cosine terms are clearly supported.

The posterior median autoregressive correlation is 0.9838, with interval $(0.9775, 0.9885)$. The temporally correlated and independent standard deviations are estimated as 0.183, with interval $(0.167, 0.201)$, and 0.105, with interval $(0.058, 0.158)$. These estimates imply the following temporal correlations:

separation	2.5%	50%	97.5%
7 days	0.853	0.892	0.922
30 days	0.505	0.613	0.706
120 days	0.065	0.142	0.248

Dependence is strong over a week, moderate over a month, and small over four months.

The posterior mean of the smooth temporal effect and its pointwise 95% credible band are shown in Figure 16. The long series contains substantial temporal structure that is not explained by the fixed covariates alone.

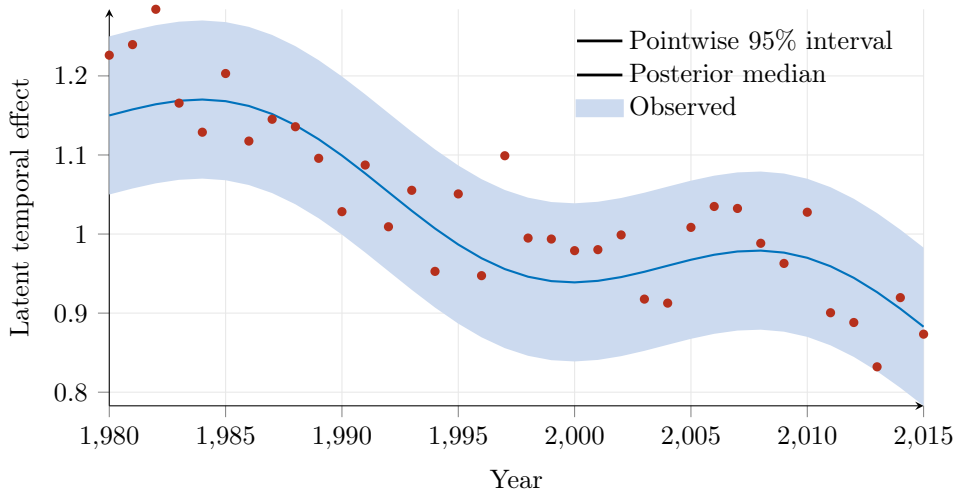

Figure 16

Integrated nested Laplace approximation returns highly accurate marginal posteriors quickly for many latent Gaussian models, but it is not universal. It does not directly return an arbitrary joint posterior, it does not support discrete latent states in the same way, the predictor must be linear on the transformed scale, and the nonlinear hyperparameter dimension must remain

modest. Questions depending on joint events, such as which day attained the largest latent risk, may require posterior sampling from an approximation or a different computational method.

Exercise 4.6 Fit a Bayesian Weibull model to the cricketer data and report a posterior credible interval for the Weibull shape parameter. Explain how the result changes under at least one defensible alternative prior.

Lecture 5 — Wednesday, February 06, 2019

Prior Specification and Model Criticism

The Fiji Fertility Survey records the number of children born to each woman, her duration of marriage, and demographic characteristics including ethnicity, literacy, residence, and age at marriage. If Y_i is the number of children and O_i is the number of years married, a baseline model is

$$Y_i \sim \text{Poisson}(O_i \lambda_i) \qquad \text{and} \qquad \log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{5.1}$$

The logarithm of exposure enters with coefficient one as an offset. Consequently, λ_i is an annual birth rate rather than an expected count over an arbitrary observation period.

The prior must be read on the scale on which it is imposed. One useful specification assigns the intercept a $\mathcal{N}(0, 10^2)$ prior and each coefficient other than the intercept an independent $\mathcal{N}(0, 0.2^2)$ prior. On the rate ratio scale, a prior standard deviation of 0.2 says that a two standard deviation change in a coefficient usually multiplies the rate by no more than

$$\exp(0.4) \approx 1.49$$

in either direction. This is weakly informative for ordinary demographic effects while suppressing implausibly enormous rate ratios. It also produces visible shrinkage for sparse categories: the estimate for a category with little information is pulled more strongly toward the reference category than an estimate supported by many observations.

Prior and posterior densities for representative Fiji coefficients are shown in Figure 17. The intercept is informed sharply by the large sample, whereas uncommon ethnic and age at marriage categories retain more prior influence. A prior should therefore be inspected coefficient by coefficient, not described globally as merely “vague” or “informative.”

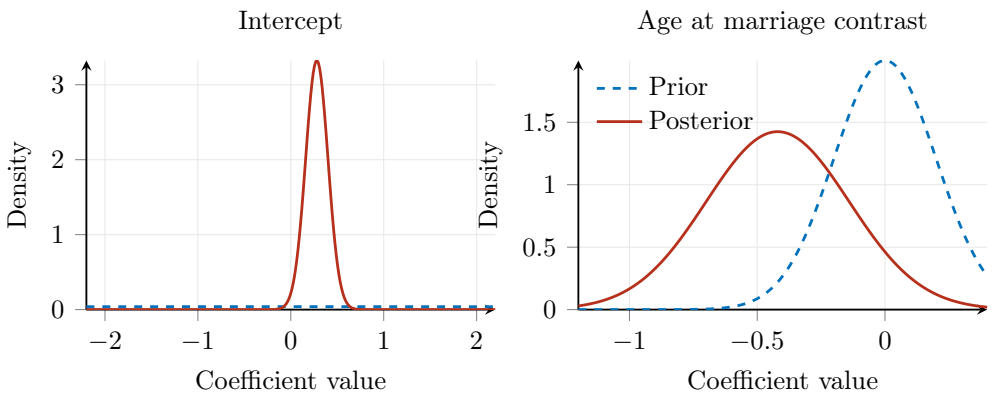

Figure 17

Count data often exhibit more zeros than a Poisson model can accommodate. A **zero-inflated Poisson model** introduces a latent structural zero indicator Z_i:

$$Z_i \sim \text{Bernoulli}(\theta) \qquad \text{and} \qquad Y_i \mid Z_i = \begin{cases} 0, & Z_i = 1, \\ \text{Poisson}(O_i\lambda_i), & Z_i = 0. \end{cases} \tag{5.2}$$

Here $Z_i = 1$ may represent infertility or another mechanism that makes a birth impossible during the observation period. A woman in the Poisson component can nevertheless have zero children. Thus

$$\mathbb{P}(Y_i = 0) = \theta + (1-\theta)e^{-O_i\lambda_i}, \tag{5.3}$$

and, for every positive integer y,

$$\mathbb{P}(Y_i = y) = (1-\theta)\frac{(O_i\lambda_i)^y e^{-O_i\lambda_i}}{y!}.$$

The parameter θ is not the observed fraction of zeros. Equation (5.3) separates structural zeros from ordinary Poisson zeros.

A beta prior on θ is especially transparent. Suppose substantive knowledge suggests a prior median of 0.025 and a prior 95% quantile of 0.10. Choosing a and b to solve

$$F^{-1}_{\text{Beta}(a,b)}(0.5) = 0.025 \qquad \text{and} \qquad F^{-1}_{\text{Beta}(a,b)}(0.95) = 0.10$$

gives approximately $a = 1.053$ and $b = 29.366$. Numerical optimization is justified here because the elicited quantiles are meaningful, while the beta shape parameters are not. Software may internally place this beta distribution on the probability obtained by applying the inverse logit transformation to an unconstrained hyperparameter.

The fitted structural zero probability is approximately 0.022, with 95% credible interval $(0.017, 0.027)$. Its posterior mode lies near a region in which numerical approximation can be delicate. A fitted model should be treated as suspicious when its answer changes materially under any of the following checks:

1. The optimizer is restarted from several plausible initial values.
2. The Gaussian, simplified Laplace, and full Laplace strategies are compared.
3. The integration grid or accuracy controls are tightened.
4. The prior is replaced by a defensible sensitivity alternative.
5. **Posterior predictive checks** compare the simulated zero counts and upper tails with those observed.

Agreement across numerical strategies does not establish that the sampling model is adequate. It only reduces concern that the approximation has found a poor mode or missed important posterior mass.

Zero inflation addresses the frequency of zeros but not necessarily extra variation among positive counts. Introduce a positive individual rate V_i satisfying

$$\mathbb{E}[V_i] = \lambda_i \qquad \text{and} \qquad \operatorname{Var}(V_i) = \frac{\lambda_i^2}{k},$$

which is obtained from a gamma distribution with shape k and rate k/λ_i. Conditional on V_i, let a nonstructural response follow $\text{Poisson}(O_i V_i)$. Integrating out V_i produces a **negative binomial distribution** with

$$\mathbb{E}[Y_i \mid Z_i = 0] = O_i\lambda_i \qquad \text{and} \qquad \operatorname{Var}(Y_i \mid Z_i = 0) = O_i\lambda_i + \frac{(O_i\lambda_i)^2}{k}. \tag{5.4}$$

Together with (5.2), this is a **zero-inflated negative binomial model**. The coefficient of variation of the latent rate is

$$\frac{\sqrt{\operatorname{Var}(V_i)}}{\mathbb{E}[V_i]} = \frac{1}{\sqrt{k}}.$$

This scale makes prior calibration easier. The Poisson base model corresponds to $1/\sqrt{k} = 0$, and a penalized complexity prior can shrink toward it. For example, the statement

$$\mathbb{P}\left(\frac{1}{\sqrt{k}} > \log(1.2)\right) = 0.5$$

uses a 20% multiplicative fluctuation as the prior median departure from the simpler model.

The fitted Fiji coefficient of variation has posterior median approximately 0.018 and 95% credible interval roughly $(0.003, 0.049)$. The structural zero probability remains near 0.022. Although the extra-Poisson component is small, the upper posterior tail permits a coefficient of variation of about 5% in the latent rates, so retaining it avoids an unjustified assertion of exact Poisson variation. The prior and posterior densities in Figure 18 show that the two nonlinear parameters occupy very different scales.

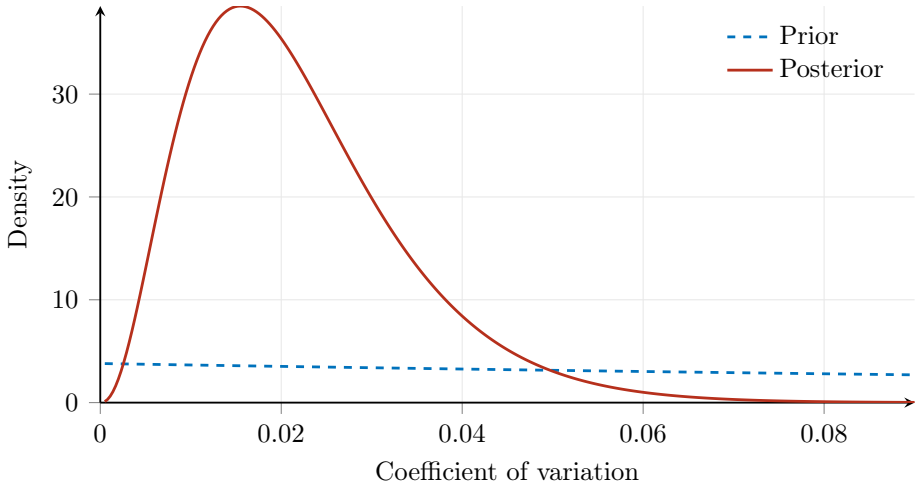

Figure 18

Regression coefficients should be reported on a scale that answers the scientific question. Exponentiating a log rate coefficient gives a rate ratio. In the Fiji analysis, the posterior mean rate ratio for rural residence relative to Suva is about 1.20, with credible interval $(1.14, 1.26)$. The corresponding value for other urban residence is 1.10, with interval $(1.04, 1.17)$. The results provide much clearer information in this form than as changes in log rate.

Model criticism should also examine the interpretation of each latent mechanism. A structural zero model asserts that some individuals have no chance of an event; a **hurdle model** instead separates zero from positive counts; a negative binomial model permits heterogeneity without declaring a distinct zero generating population. Similar likelihood values do not make these scientific stories interchangeable.

Posterior Prediction and Joint Inference

The Gaussian approximation used by integrated nested Laplace approximation is most transparent when the hyperparameters are fixed. Write the bacteria model in vector form as

$$Y_k \mid \eta_k \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_k)) \qquad \text{and} \qquad \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{U},$$

where row k of $\mathbf{A}$ identifies the individual responsible for observation k. Conditional on a fixed random effect standard deviation σ, the latent field

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{U} \\ \boldsymbol{\beta} \end{pmatrix}$$

has a Gaussian prior with sparse precision $\mathbf{Q}_0(\sigma)$. The linear constraint for $\boldsymbol{\eta}$ can be represented by an auxiliary Gaussian distribution with variance $\epsilon\mathbf{I}$, followed by the limit $\epsilon \downarrow 0$.

Let $\widehat{\boldsymbol{W}}(\sigma)$ maximize the conditional log posterior. Conditional independence makes the negative Hessian of the log likelihood diagonal in the predictor coordinates. If this matrix is denoted by $\mathbf{C}(\widehat{\boldsymbol{W}}(\sigma))$, then the Gaussian approximation is

$$\widetilde{\pi}_G(\boldsymbol{W} \mid \boldsymbol{y}, \sigma) = \mathcal{N}\left(\widehat{\boldsymbol{W}}(\sigma), \left\{\mathbf{Q}_0(\sigma) + \mathbf{C}(\widehat{\boldsymbol{W}}(\sigma))\right\}^{-1}\right). \tag{5.5}$$

The approximation uses sparse factorization rather than forming the inverse in (5.5). Its marginal means and variances are inexpensive, but skewness in binary response posteriors can make a purely Gaussian approximation inaccurate.

For a particular component $W_j = w$, the fuller Laplace approximation is based on the identity, up to a normalizing constant that does not depend on w,

$$\pi(W_j = w \mid \boldsymbol{y}, \sigma) \propto \frac{\pi(\boldsymbol{W}, \boldsymbol{y} \mid \sigma)}{\pi(\boldsymbol{W}_{-j} \mid W_j = w, \boldsymbol{y}, \sigma)}.$$

The numerator is evaluated at the constrained mode

$$\widehat{\boldsymbol{W}}_{-j}(w, \sigma) = \arg\max_{\boldsymbol{W}_{-j}} \pi(\boldsymbol{W}_{-j}, W_j = w \mid \boldsymbol{y}, \sigma),$$

and a Gaussian approximation replaces the denominator. Recomputing this constrained mode over a grid of w-values is accurate but costly. The **simplified Laplace method** avoids full reoptimization and uses higher derivatives to correct the Gaussian approximation for skewness.

If σ is unknown, an **empirical Bayes calculation** plugs in the mode of $\pi(\sigma \mid \boldsymbol{y})$. Full integrated nested Laplace approximation instead evaluates a collection of support points $\sigma_1, \ldots, \sigma_G$ spanning the important posterior mass and computes

$$\widetilde{\pi}(W_j \mid \boldsymbol{y}) = \sum_{g=1}^{G} \widetilde{\pi}(W_j \mid \boldsymbol{y}, \sigma_g)\widetilde{\pi}(\sigma_g \mid \boldsymbol{y})\Delta_g. \tag{5.6}$$

For a vector of several hyperparameters, the support points form a multidimensional design. The number of required evaluations increases rapidly with dimension, which explains why the hyperparameter vector, rather than the latent field, is often the computational bottleneck.

In the bacteria example with $\sigma = 1$, a Gaussian approximation gives a 95% interval $(0.041, 1.680)$ for the placebo indicator, while the full Laplace interval is $(0.135, 1.784)$. The simplified Laplace interval $(0.127, 1.766)$ is very close to the full calculation. The comparison illustrates why a skewness correction matters even when posterior means appear stable.

Prediction begins by stating the target. For a new design matrix $\widetilde{\mathbf{X}}$, a likelihood analysis approximates the distribution of the new linear predictors by

$$\widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}} \dot{\sim} \mathcal{N}\left(\widetilde{\mathbf{X}}\boldsymbol{\beta}, \widetilde{\mathbf{X}}\mathcal{I}(\widehat{\boldsymbol{\beta}})^{-1}\widetilde{\mathbf{X}}^{\top}\right). \tag{5.7}$$

For a Bayesian analysis, each row $\widetilde{\boldsymbol{x}}_r^{\top}$ defines a linear combination $\widetilde{\boldsymbol{x}}_r^{\top}\boldsymbol{\beta}$. Its marginal posterior can be computed directly because it is part of the latent Gaussian structure.

For the Fiji model, one useful target is the expected number of children per decade among women in selected ethnicity and age at marriage groups, with literacy and residence held at reference values. A compact R fragment creates the new design and registers its rows as linear combinations:

```
new_data <- expand.grid(
  ageMarried = levels(fiji$ageMarried),
  ethnicity = levels(fiji$ethnicity)
)
new_matrix <- model.matrix(~ ethnicity * ageMarried, new_data)
new_combinations <- INLA::inla.make.lincombs(as.data.frame(new_matrix))
```

Only the construction of the inferential target is retained here; routine fitting and plotting commands do not add statistical content.

The resulting posterior medians and pointwise 95% credible bands show how the estimated birth rate changes with age at marriage and ethnicity. They refer to the latent mean for a fertile population under the chosen reference conditions.

The posterior birth rate curves across marriage age and ethnicity groups are shown in Figure 19.

Three distinct predictive quantities must not be conflated. Under the zero-inflated negative bino-

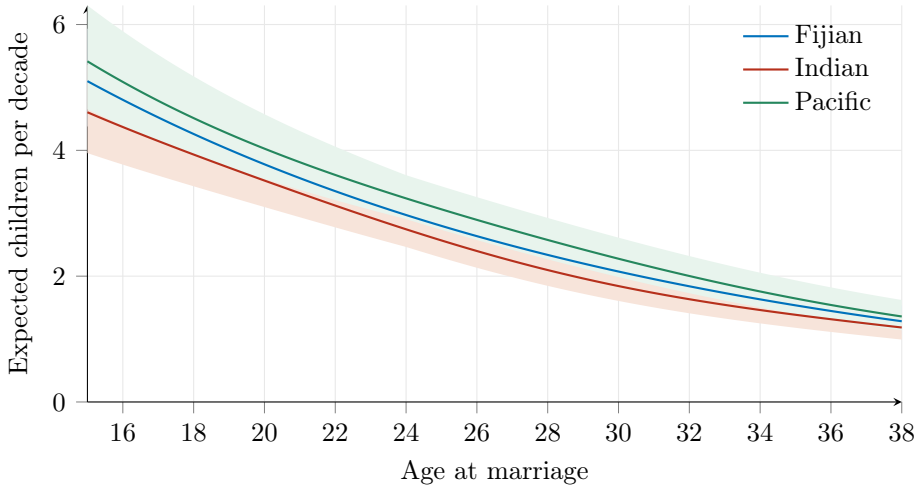

Figure 19

mial hierarchy,

1. The expected count for the fertile component is $O_i \lambda_i$.
2. The population mean, including structural zeros, is

$$\mathbb{E}[Y_i \mid \boldsymbol{\beta}, \theta] = (1 - \theta) O_i \lambda_i.$$

3. A future individual's count additionally varies through the gamma effect and the Poisson sampling distribution.

A credible interval for a mean function is therefore generally narrower than a **posterior predictive interval** for a new individual. The distinction should be stated whenever a plot is described as a prediction.

Marginal posterior summaries are insufficient for a nonlinear function involving several parameters. For example,

$$M_r = 10(1 - \theta) \exp\left(\widetilde{\boldsymbol{x}}_r^\top \boldsymbol{\beta}\right)$$

depends jointly on θ and $\boldsymbol{\beta}$. Drawing each coefficient independently from its marginal posterior destroys posterior correlations and gives the wrong uncertainty. Instead, draw a **joint approximate posterior sample**

$$(\boldsymbol{\beta}^{(s)}, \theta^{(s)}, k^{(s)}), \quad s = 1, \ldots, S,$$

and compute $M_r^{(s)}$ for every draw. Quantiles of the derived values then approximate the posterior

of M_r.

If individual heterogeneity is part of the target, also draw

$$V_r^{(s)} \mid \boldsymbol{\beta}^{(s)}, k^{(s)} \sim \text{Gamma}\left(\text{shape} = k^{(s)}, \text{rate} = \frac{k^{(s)}}{\exp\left(\widetilde{\boldsymbol{x}}_r^\top \boldsymbol{\beta}^{(s)}\right)}\right),$$

then include the structural zero indicator and, for full posterior prediction, a Poisson response. This staged simulation mirrors the hierarchy and makes the estimand explicit.

Posterior draws of the gamma shape and structural zero probability may be strongly dependent, as Figure 20 illustrates. Their joint geometry is precisely what independent marginal sampling would discard.

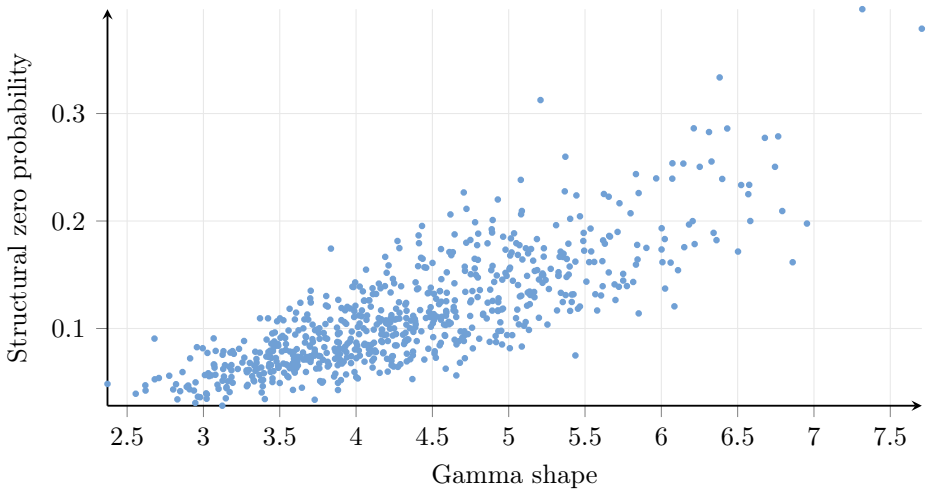

Figure 20

Clear reporting combines these elements: priors are justified on interpretable scales, nonlinear parameters are transformed before presentation, coefficient tables use rate ratios when appropriate, and predicted values supplement rather than replace parameter estimates. Numerical approximation, prior sensitivity, sampling model adequacy, and predictive uncertainty are separate questions and should be assessed separately.

Lecture 6 — Wednesday, February 13, 2019

4. Survival, Dependence, and Semiparametric Models

Survival Distributions, Hazards, and Censoring

Survival analysis concerns a nonnegative event time T, such as lifetime, time to infection, or age at initiation of a behaviour. The central complication is that the event time is often only partially observed.

Definition 6.1 If T has distribution function F and density f, its **survival function**, **hazard function**, and **cumulative hazard function** are

$$S(t) = \mathbb{P}(T > t) = 1 - F(t),$$

$$h(t) = \lim_{\delta \downarrow 0} \frac{\mathbb{P}(t \le T < t + \delta \mid T \ge t)}{\delta} = \frac{f(t)}{S(t)},$$

and

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = -\log S(t),$$

respectively.

The hazard is an instantaneous event rate conditional on survival to time t; it is not a probability. The identities in Definition 6.1 imply

$$S(t) = e^{-H(t)} \qquad \text{and} \qquad f(t) = h(t)e^{-H(t)}.$$

Thus any one of f, S, h, and H determines the others.

Definition 6.2 A random variable T has a **Weibull distribution** with scale $\lambda > 0$ and shape $\kappa > 0$ if

$$f(t; \lambda, \kappa) = \frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa - 1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\kappa}\right), \quad t > 0.$$

For the Weibull distribution,

$$S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\kappa}\right), \qquad h(t) = \frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa-1}, \qquad \text{and} \qquad H(t) = \left(\frac{t}{\lambda}\right)^{\kappa}. \tag{6.1}$$

Its moments are

$$\mathbb{E}[T] = \lambda\Gamma\left(1 + \frac{1}{\kappa}\right)$$

and

$$\mathrm{Var}(T) = \lambda^2\left[\Gamma\left(1 + \frac{2}{\kappa}\right) - \Gamma\left(1 + \frac{1}{\kappa}\right)^2\right].$$

When $\kappa = 1$, the hazard is constant and the distribution is exponential. The hazard increases when $\kappa > 1$ and decreases when $0 < \kappa < 1$. The effects of changing the scale and shape are shown in Figure 21.

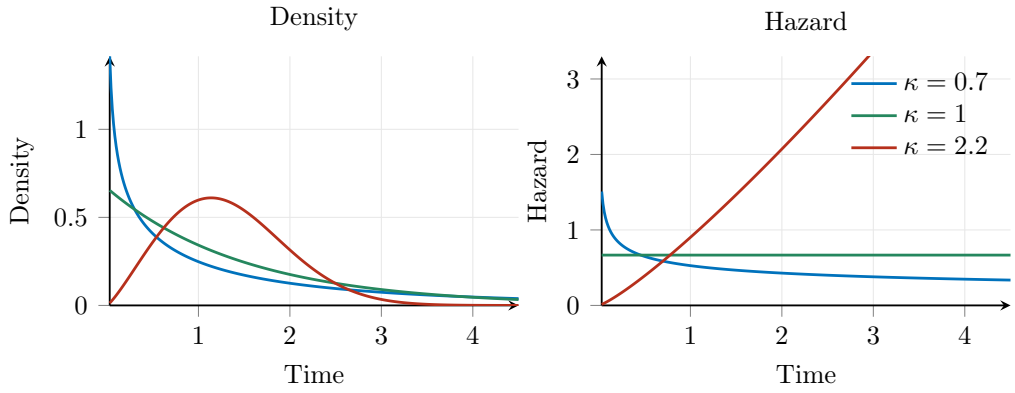

Figure 21

There are two common regression parameterizations. An **accelerated failure time model** uses

$$\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\gamma}.$$

Increasing a coefficient stretches or contracts the time axis: if two individuals differ by one unit in covariate q, their corresponding quantiles differ by the time ratio e^{γ_q}. A **proportional hazards model** instead writes

$$h_i(t) = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} h_0(t),$$

so e^{β_q} is a hazard ratio.

The Weibull family supports both representations. If

$$\rho_i = \lambda_i^{-\kappa} \qquad \text{and} \qquad \log(\rho_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

then

$$h_i(t) = \rho_i \kappa t^{\kappa-1}.$$

With common shape, $\boldsymbol{\beta} = -\kappa\boldsymbol{\gamma}$. The diagrams in Figure 22 show how the two parameterizations alter the hazard.

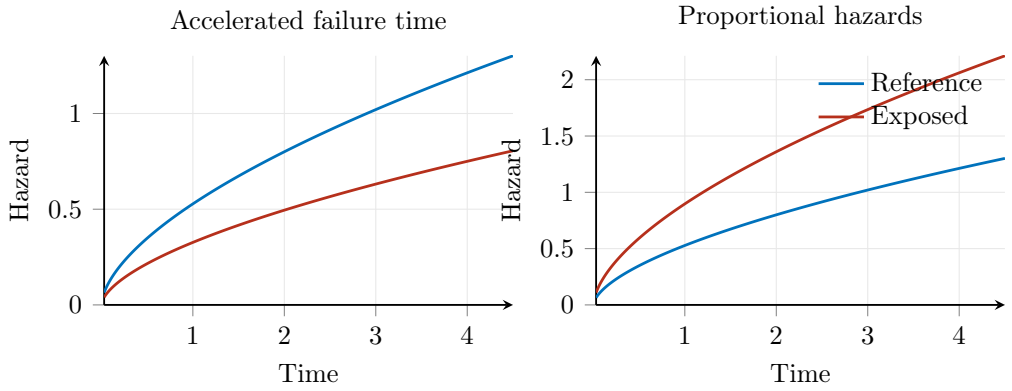

Figure 22

Software conventions must therefore be checked rather than guessed. One useful convention sets

$$-\log(\lambda_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{6.2}$$

Then e^{β_q} is a ratio of reciprocal scales, or event speeds. A positive coefficient shortens the time scale by the factor $e^{-\beta_q}$. It is not the proportional hazard ratio unless the coefficient is multiplied by κ.

The cricketer data provide a representative example. Lifetime is regressed on birth decade and left-handedness. For players born before 1890, a likelihood fit gives Weibull shape approximately 5.35. The left-handedness coefficient is close to zero, while birth decade is associated with longer lifetimes. A Bayesian fit uses a normal prior for $\log \kappa$ with mean $\log(7.5)$ and standard deviation $2/3$; equivalently, the prior median of κ is 7.5. Its central prior interval is approximately $(2.03, 27.70)$. This prior permits a broad range of left-skewed lifetime densities.

The fitted density and cumulative hazard agree reasonably with their empirical counterparts. Posterior draws, rather than a single curve at the posterior mode, display uncertainty in both the density and cumulative hazard.

The fitted density and cumulative hazard are compared with their empirical counterparts in Figure 23.

If every event time were observed exactly, the likelihood would be $\prod_i f(t_i)$. More commonly, **right**

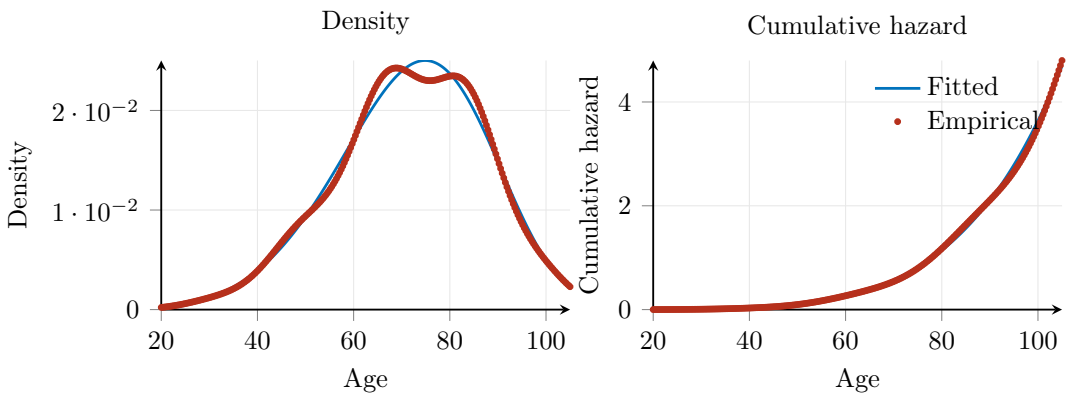

Figure 23

censoring occurs when observation ends at a censoring time C_i. Define

$$Z_i = \min(T_i, C_i) \qquad \text{and} \qquad \Delta_i = \mathbb{1}\{T_i \leq C_i\}.$$

Conditional on the explanatory variables, assume that T_i and C_i are independent. The observed data likelihood contribution is then

$$L_i = f(Z_i)^{\Delta_i} S(Z_i)^{1-\Delta_i}. \tag{6.3}$$

An exact event contributes its density. A right censored observation contributes the probability of surviving beyond its last observed time. **Informative censoring** violates the conditional independence assumption and cannot be repaired merely by using (6.3).

Including living cricketers as right censored observations yields posterior Weibull shape 5.641, with 95% credible interval $(5.489, 5.796)$. Under the event speed convention in (6.2), the left-handedness ratio is 1.006, with interval $(0.990, 1.021)$. Left-handed players therefore have an estimated time scale about 0.6% shorter, with an interval ranging from approximately 2.1% shorter to 1.0% longer. A birth year one century later has event speed ratio 0.769, with interval $(0.749, 0.790)$; on the time scale, this is approximately a 2.6% increase per decade.

Interval censoring is handled by the same principle. If an event is known only to lie in $(L_i, R_i]$, its contribution is

$$\mathbb{P}(L_i < T_i \leq R_i) = F(R_i) - F(L_i) = S(L_i) - S(R_i). \tag{6.4}$$

Exact observations arise as a density limit, right censoring sets $R_i = \infty$, and left censoring sets $L_i = 0$. Consequently, a mixed collection of exact, right censored, left censored, and interval

censored data has likelihood

$$L(\boldsymbol{\psi}) = \prod_{i \in \mathcal{E}} f(t_i; \boldsymbol{\psi}) \prod_{i \in \mathcal{R}} S(L_i; \boldsymbol{\psi}) \prod_{i \in \mathcal{L}} F(R_i; \boldsymbol{\psi}) \prod_{i \in \mathcal{I}} \{F(R_i; \boldsymbol{\psi}) - F(L_i; \boldsymbol{\psi})\}.$$

Scaling time so that its values are numerically moderate changes the intercept but not the scientific model. It can greatly improve optimization and numerical differentiation.

Hierarchical Survival and Clustered Binary Models

Repeated event times from the same subject are dependent. A shared random effect, traditionally called a **frailty**, represents this dependence. Under the event speed parameterization,

$$T_{ij} \mid U_i \sim \text{Weibull}(\lambda_{ij}, \kappa), \qquad -\log(\lambda_{ij}) = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + U_i, \qquad \text{and} \qquad U_i \sim \mathcal{N}(0, \sigma^2). \tag{6.5}$$

Conditional on U_i, event times are independent. Marginally, a positive U_i gives all observations from subject i shorter time scales, creating positive association within a subject.

The kidney data contain times to infection for 38 dialysis patients, with two observations per patient. The explanatory variables are age, sex, and disease type. A prior

$$\log \kappa \sim \mathcal{N}(0, 1)$$

has a central interval approximately $(0.141, 7.10)$. For the frailty standard deviation, the calibration

$$\mathbb{P}(\sigma > 0.55) = 0.05$$

states that, with high prior probability, the central 95% of event speed multipliers specific to each subject lie between approximately $1/3$ and 3, since

$$\exp(\pm 2 \times 0.55) \approx (0.333, 3.00).$$

The posterior median frailty standard deviation is 0.24, with 95% credible interval $(0.05, 0.56)$. The posterior median shape is 0.98, with interval $(0.74, 1.17)$, so an exponential baseline hazard remains plausible. The female coefficient is -1.61, with interval $(-2.28, -0.91)$, indicating a substantially lower event speed. The coefficient for polycystic kidney disease is -1.23, with interval $(-2.48, 0.02)$, suggesting a potentially important but uncertain reduction relative to the reference disease.

Only two observations are available per patient, so the likelihood contains limited information about σ. The prior and posterior in Figure 24 show that the frailty variance is supported away from zero but still has a substantial upper tail. A prior allowing greater variation would give more

conservative fixed effect inference, whereas forcing σ to be small would understate dependence within a patient.

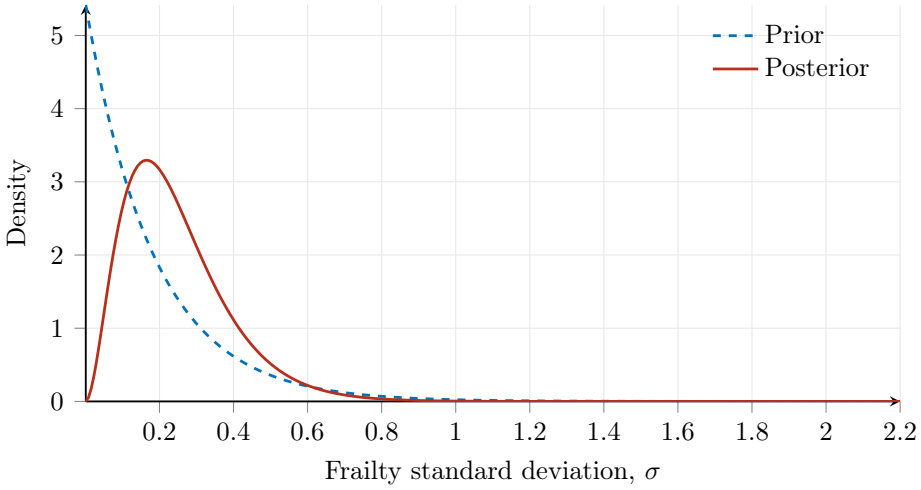

Figure 24

The Fiji survey illustrates interval censoring. Age at first marriage is recorded in categories such as 15–18, 18–20, and 20–22, so the exact event time is unknown. Women not yet married are right censored at their current age. Taking age 12 as the time origin, the response for a married woman is an interval $(L_i, R_i]$, and the response for an unmarried woman is (L_i, ∞).

A Weibull model with ethnicity and residence gives shape approximately 1.55. Relative to Fijian women, the estimated event speed ratios are 1.31 for Indian women, 2.03 for women of part European ancestry, 1.51 for Pacific Islander women, 1.81 for Rotuman women, and 1.83 for Chinese women. An event speed ratio greater than one corresponds to an earlier age at marriage under this convention.

The fitted densities in Figure 25 compare selected ethnic groups. The distribution for Indian women is shifted toward younger ages relative to that for Fijian women, while the model also propagates uncertainty in the Weibull shape and regression coefficients.

The extremely narrow numerical posterior reported for the shape parameter is a reason for scrutiny, not automatically evidence of precision. With a large sample, concentration may be genuine, but it should be checked against alternative starting values, integration strategies, time scaling, and direct likelihood calculations. A model can be scientifically inadequate even when its numerical posterior is stable.

Age at first cigar use provides a second interval censored example with a deeper hierarchy. If

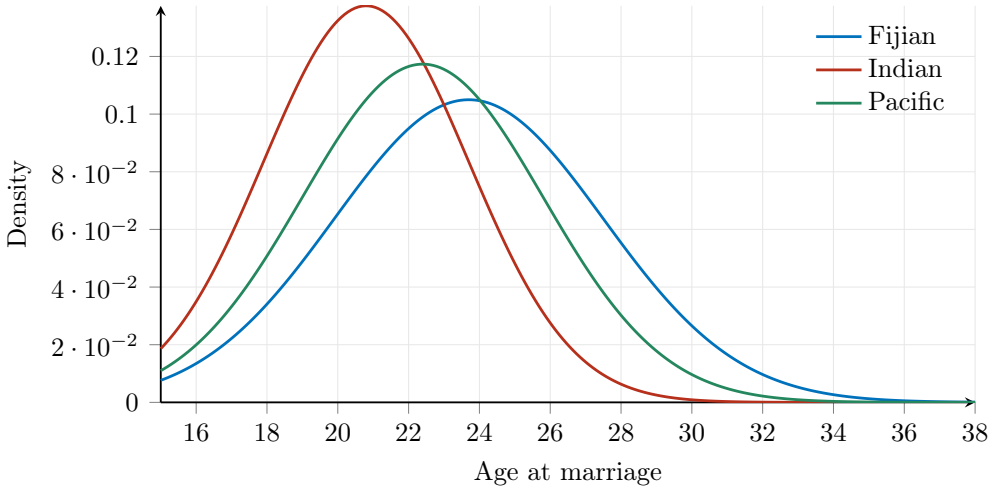

Figure 25

initiation is reported at age a, the event is known to occur in $[a, a+1)$; respondents who have not initiated are right censored at their current age. A hierarchy with two levels is

$$T_{ijk} \mid U_i, V_{ij} \sim \text{Weibull}(\lambda_{ijk}, \kappa), \tag{6.6}$$

$$-\log(\lambda_{ijk}) = \boldsymbol{x}_{ijk}^\top \boldsymbol{\beta} + U_i + V_{ij}, \qquad U_i \sim \mathcal{N}(0, \sigma_U^2), \qquad \text{and} \qquad V_{ij} \sim \mathcal{N}(0, \sigma_V^2),$$

where i indexes states, j indexes schools within states, and k indexes students.

The posterior median Weibull shape is approximately 7.96, with interval $(7.74, 8.19)$. The school standard deviation is 0.062, with interval $(0.051, 0.074)$, and the state standard deviation is 0.031, with interval $(0.015, 0.054)$. Both clustering levels contribute, although variation among schools is larger. Among the fixed effects, the coefficient for Asian respondents is -0.147, with interval $(-0.186, -0.111)$, and the female coefficient is -0.057, with interval $(-0.066, -0.048)$.

The prior and posterior distributions for the shape and the two standard deviations are shown in Figure 26. Each variance component is calibrated separately because its grouping level has a different interpretation.

Hierarchical survival models demand care at three levels. First, the Weibull parameterization determines whether a coefficient is a time ratio, reciprocal scale ratio, or hazard ratio. Second, censoring determines the appropriate likelihood contribution. Third, each frailty distribution and prior determines how dependence and heterogeneity are represented. Writing the complete probability model, including the observation mechanism, is the most reliable way to keep these interpretations aligned.

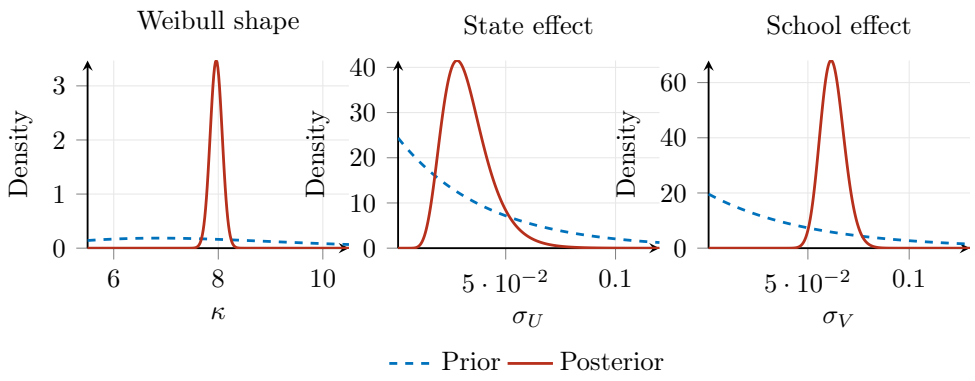

Figure 26

Lecture 7 — Wednesday, February 27, 2019

Temporal Dependence and Correlated Random Effects

A random intercept gives every pair of observations from the same individual the same covariance, regardless of their separation in time. This is unsuitable when measurements one day apart should be more similar than measurements ten days apart. A **temporal process** replaces exchangeable correlation by a function of the time lag.

Definition 7.1 A stationary **first order autoregressive process** satisfies

$$V_{t+1} = \theta V_t + \varepsilon_{t+1} \qquad \text{and} \qquad \varepsilon_t \sim \mathcal{N}(0, \nu^2)$$

independently, where $|\theta| < 1$, and

$$V_1 \sim \mathcal{N}\left(0, \frac{\nu^2}{1-\theta^2}\right).$$

Iterating the recursion in Definition 7.1 gives

$$V_{t+k} = \theta^k V_t + \sum_{r=1}^{k} \theta^{k-r} \varepsilon_{t+r}. \tag{7.1}$$

Therefore,

$$\mathbb{E}[V_{t+k} \mid V_t] = \theta^k V_t$$

and

$$\mathrm{Var}(V_{t+k} \mid V_t) = \nu^2 \sum_{r=0}^{k-1} \theta^{2r} = \nu^2 \frac{1-\theta^{2k}}{1-\theta^2}.$$

The stationary covariance and correlation are

$$\mathrm{Cov}(V_t, V_{t+k}) = \frac{\nu^2}{1-\theta^2}\theta^{|k|} \qquad \text{and} \qquad \mathrm{Corr}(V_t, V_{t+k}) = \theta^{|k|}. \tag{7.2}$$

Large positive θ produces smooth, persistent series; values near zero produce nearly white noise. The contrast is visible in Figure 27.

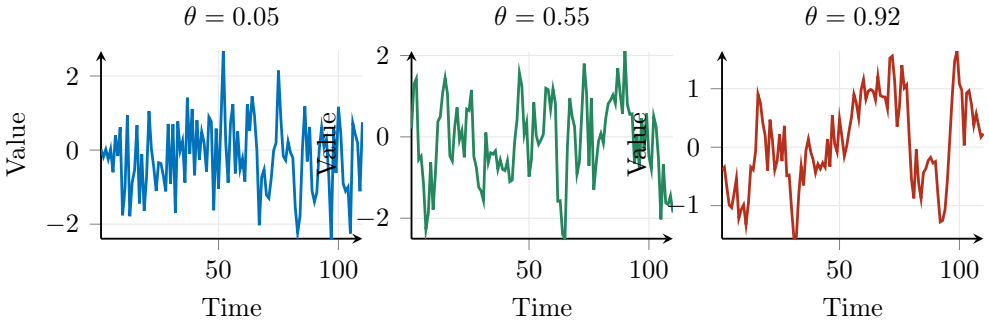

Figure 27

At $\theta = 1$, the recursion becomes a random walk and is no longer stationary. Negative θ produces alternating correlation and is possible in discrete time. A continuous time process with exponential correlation cannot reproduce that alternation.

The continuous time analogue is the **Ornstein–Uhlenbeck process**,

$$\mathrm{d}V(t) = -\alpha V(t)\,\mathrm{d}t + \gamma\,\mathrm{d}B(t), \quad \alpha > 0, \tag{7.3}$$

where $B(t)$ is a standard Brownian motion. In stationarity,

$$\mathbb{E}[V(t+h) \mid V(t)] = e^{-\alpha h} V(t), \quad h \geq 0,$$

and

$$\mathrm{Cov}(V(t+h), V(t)) = \sigma^2 e^{-|h|/\phi}, \qquad \sigma^2 = \frac{\gamma^2}{2\alpha}, \qquad \text{and} \qquad \phi = \frac{1}{\alpha}.$$

For observations one time unit apart, $\theta = e^{-\alpha} = e^{-1/\phi}$. More generally, the conditional innovation

variance over lag h is

$$\mathrm{Var}(V(t+h) \mid V(t)) = \sigma^2\{1 - e^{-2h/\phi}\}.$$

The **range** ϕ has units of time, while σ^2 is the stationary contribution of the process and does not depend on whether time is measured in days or weeks.

Definition 7.2 For a weakly stationary process with covariance $C(h)$, the **semivariogram** is

$$\gamma(h) = \frac{1}{2}\mathbb{E}\left[\{V(t+h) - V(t)\}^2\right] = C(0) - C(h).$$

If $C(h) = \sigma^2\rho(h)$, then

$$\gamma(h) = \sigma^2\{1 - \rho(h)\}.$$

The covariance and semivariogram therefore contain the same second order information. For exponential correlation, the covariance decreases from σ^2 to zero and the semivariogram increases from zero to the **sill** σ^2. The range controls how quickly both transitions occur.

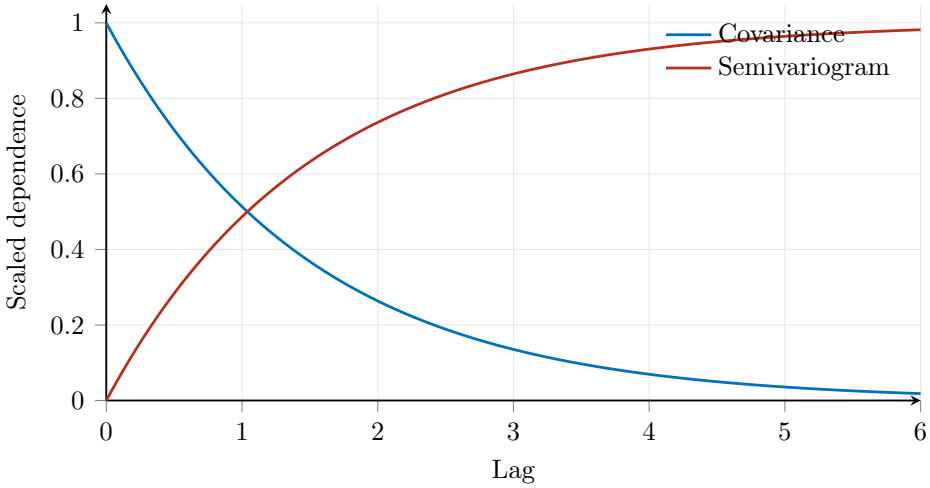

Figure 28

The covariance and variogram encode the same dependence pattern in the two panels of Figure 28.

The semivariogram is particularly useful with irregularly spaced observations because pairs can be grouped by their actual separation rather than by a common discrete lag. This same idea will transfer directly to spatial separation.

A **longitudinal mixed model** can contain both persistent individual heterogeneity and dependence that varies with time:

$$Y_{ij} = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + U_i + V_i(t_{ij}) + \varepsilon_{ij}, \tag{7.4}$$

where

$$U_i \sim \mathcal{N}(0, \sigma_U^2), \qquad \mathrm{Cov}(V_i(t+h), V_i(t)) = \sigma_V^2 e^{-|h|/\phi}, \qquad \text{and} \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \tau^2).$$

The processes V_i are independent across individuals and independent of the random intercepts and errors. It follows that

$$\mathrm{Var}(Y_{ij}) = \sigma_U^2 + \sigma_V^2 + \tau^2,$$

and

$$\mathrm{Cov}(Y_{ij}, Y_{ik}) = \sigma_U^2 + \sigma_V^2 \exp\left(-\frac{|t_{ij} - t_{ik}|}{\phi}\right) + \tau^2 \mathbb{1}\{j = k\}. \tag{7.5}$$

Responses from different individuals are uncorrelated under this model.

Let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iJ_i})^\top$, and define

$$\{\mathbf{R}_i(\phi)\}_{jk} = \exp\left(-\frac{|t_{ij} - t_{ik}|}{\phi}\right).$$

The marginal model has block diagonal covariance with blocks specific to each subject

$$\boldsymbol{\Gamma}_i = \sigma_U^2 \mathbf{1}_{J_i} \mathbf{1}_{J_i}^\top + \sigma_V^2 \mathbf{R}_i(\phi) + \tau^2 \mathbf{I}_{J_i}. \tag{7.6}$$

Likelihood and restricted likelihood calculations proceed exactly as for the mixed models developed earlier, with $\boldsymbol{\Gamma}_i$ replacing the exchangeable covariance block.

The four covariance parameters play different roles. The random intercept controls correlation that persists at arbitrarily long lags. The temporal variance controls the magnitude of the decaying component, the range controls its duration, and the **nugget** τ^2 represents measurement error or variation at a scale shorter than the sampling interval. Estimating all four reliably requires enough repeated observations to distinguish these signatures.

The pig weight data contain nine measurements on each of 48 pigs. An exponential correlation fit estimates a negligible random intercept standard deviation and a negligible nugget, temporal standard deviation 4.681, and range 17.90 measurement intervals. Because the fitted range exceeds the observed span, the temporal process changes only slowly within a pig and is difficult to distinguish from growth variation specific to each subject. The left panel of Figure 29 shows the observed trajectories, and the right panel shows fitted straight trajectories from a random intercept and random slope model.

A random intercept and slope model offers a scientifically natural alternative:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + U_{0i} + U_{1i} t_{ij} + \varepsilon_{ij}. \tag{7.7}$$

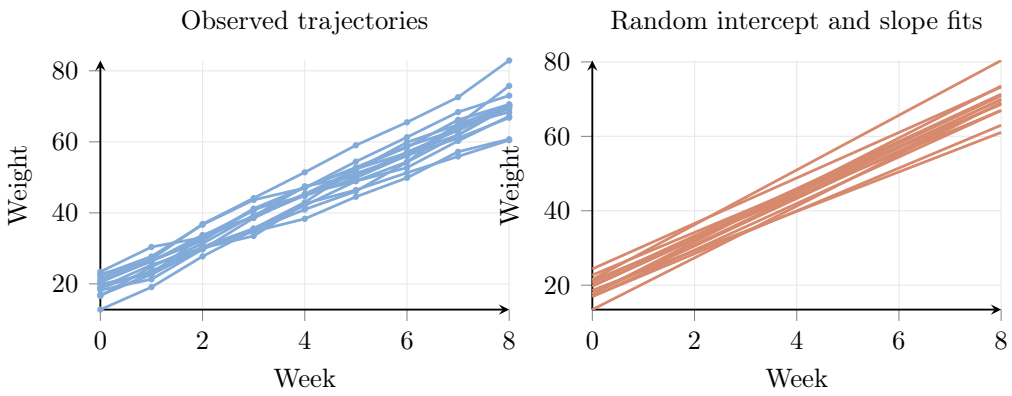

Figure 29

The fitted fixed intercept and slope are 19.356 and 6.210. The fitted standard deviations are 2.643 for the random intercept, 0.616 for the random slope, and 1.264 for the residual, with random effect correlation -0.063. Unlike the exponential temporal model, (7.7) says that each pig follows an approximately straight trajectory with its own starting weight and growth rate.

The random slope and serial correlation models are not nested. A likelihood ratio test between them is therefore not available. Selection should combine scientific mechanism, predictive performance, residual diagnostics, and sensitivity to the limited observation span. Smooth systematic divergence between subjects favours random slopes; short term fluctuations whose similarity decays with lag favour a temporal covariance model. With few measurements per individual, the data may not identify that distinction sharply.

Conditional prediction retains the usual multivariate normal form. If $\boldsymbol{W}$ collects the random intercepts and sampled temporal effects, then

$$\mathbb{E}[\boldsymbol{W} \mid \boldsymbol{Y}] = \operatorname{Cov}(\boldsymbol{W}, \boldsymbol{Y}) \operatorname{Var}(\boldsymbol{Y})^{-1} \{\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\}.$$

The only change is that the covariance matrices are constructed from (7.5) and (7.6). Prediction at an unobserved time additionally uses the covariance between that time and the observed times.

Lecture 8 — Wednesday, March 06, 2019

Generalized Additive Models and Splines

A linear predictor need not force every continuous explanatory variable to have a linear effect. A **generalized additive model** replaces selected linear terms by smooth functions:

$$Y_i \sim \mathcal{F}(\mu_i, \boldsymbol{\theta}) \qquad \text{and} \qquad g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + f_1(w_{i1}) + \cdots + f_K(w_{iK}). \tag{8.1}$$

The distribution and link retain the generalized linear model structure. The functions f_k are estimated from the data subject to smoothness assumptions.

Without a restriction, a sufficiently flexible function can interpolate every observation and has no predictive value. For one smooth term, a standard penalized log likelihood is

$$\ell_p(\boldsymbol{\beta}, f; \alpha) = \ell(\boldsymbol{\beta}, f; \boldsymbol{y}) - \frac{\alpha}{2} \int \{f''(u)\}^2 \, \mathrm{d}u. \tag{8.2}$$

The integral is a measure of curvature. The **smoothing parameter** α governs the compromise: small α follows the data closely, while large α suppresses curvature. As α tends to infinity, the fitted function approaches its unpenalized null space, which consists of straight lines for the second derivative penalty.

The smooth is not separately identifiable from the intercept unless a constraint is imposed. A common choice is

$$\sum_{i=1}^{N} f(w_i) = 0.$$

The intercept then represents the average fitted level, while f represents deviations from it.

Theorem 8.1 If a maximizer of (8.2) exists, it may be chosen to be a **natural cubic spline** with knots at the distinct observed covariate values.

The natural cubic spline is piecewise cubic, with continuous first and second derivatives at every knot, and is linear beyond the boundary knots. This result converts an infinite dimensional optimization into a finite dimensional one.

Write a spline in a basis as

$$f(w) = \boldsymbol{b}(w)^\top \boldsymbol{a},$$

and let $\mathbf{B}$ have rows $\boldsymbol{b}(w_i)^\top$. The roughness penalty becomes

$$\int \{f''(u)\}^2 \, \mathrm{d}u = \boldsymbol{a}^\top \mathbf{S} \boldsymbol{a}, \tag{8.3}$$

where $\mathbf{S}$ is positive semidefinite. Cubic B-splines provide a local basis: each **basis function** is nonzero over only a short interval, which makes $\mathbf{B}$ and the associated computations sparse. The upper panel of Figure 30 shows the basis functions, and the lower panel shows curves obtained from different coefficient vectors.

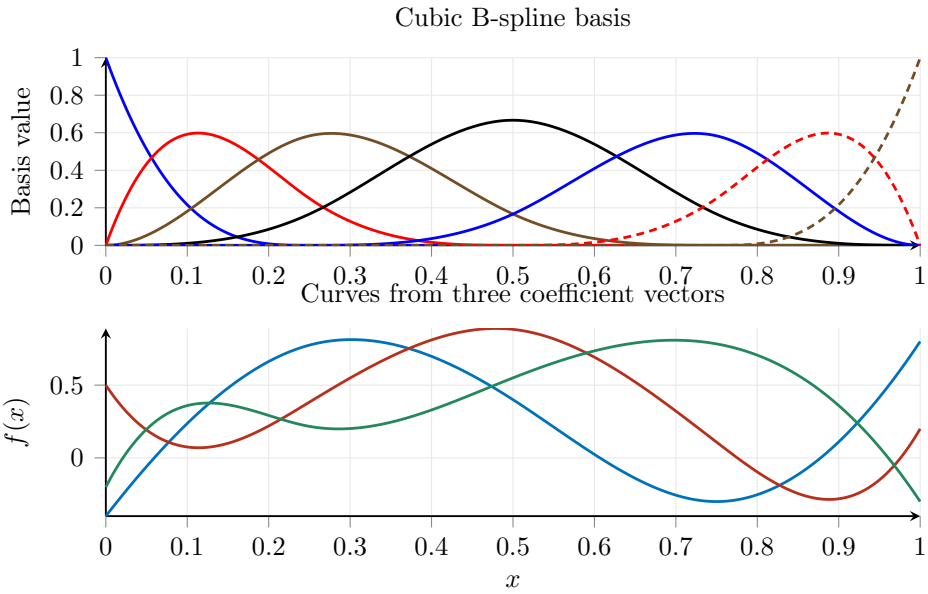

Figure 30

For Gaussian responses with working weight matrix $\mathbf{Q}$, the penalized coefficient estimate has the form

$$\widehat{\boldsymbol{a}}_\alpha = \left(\mathbf{B}^\top \mathbf{Q} \mathbf{B} + \alpha \mathbf{S}\right)^{-1} \mathbf{B}^\top \mathbf{Q} \boldsymbol{y}.$$

The corresponding influence matrix is

$$\mathbf{H}_\alpha = \mathbf{B} \left(\mathbf{B}^\top \mathbf{Q} \mathbf{B} + \alpha \mathbf{S}\right)^{-1} \mathbf{B}^\top \mathbf{Q}. \tag{8.4}$$

The quantity $\mathrm{tr}(\mathbf{H}_\alpha)$ is the **effective degrees of freedom**. It decreases as smoothing strengthens and is usually far smaller than the number of basis coefficients.

Leave one out cross validation chooses α by predictive performance, but refitting N times is expensive. For Gaussian data, **generalized cross validation** uses

$$\mathrm{GCV}(\alpha) = \frac{N\|\boldsymbol{y} - \widehat{\boldsymbol{y}}_\alpha\|^2}{\{N - \mathrm{tr}(\mathbf{H}_\alpha)\}^2}. \tag{8.5}$$

For non-Gaussian responses, an analogous criterion is evaluated during iteratively reweighted fitting. Restricted likelihood provides another principled way to estimate smoothing parameters through the random effect representation.

Indeed, (8.3) is proportional to the log density of a Gaussian coefficient vector with precision $\alpha\mathbf{S}$:

$$\pi(\boldsymbol{a} \mid \alpha) \propto \exp\left(-\frac{\alpha}{2}\boldsymbol{a}^\top \mathbf{S}\boldsymbol{a}\right).$$

Because $\mathbf{S}$ has a null space, this is an **intrinsic Gaussian prior** rather than a proper distribution until the unpenalized components are constrained or represented separately as fixed effects. Penalized smoothing, random effects, and Bayesian Gaussian priors are therefore different inferential interpretations of closely related algebra.

The mathematics achievement data relate a student's score to socioeconomic status, minority status, sex, and school. A basic model is

$$Y_{ijk} = \boldsymbol{x}_{ijk}^\top \boldsymbol{\beta} + f(w_{ijk}) + U_i + \varepsilon_{ijk},$$

where U_i is a school random intercept and w_{ijk} is socioeconomic status. Separate curves by minority status can be written

$$Y_{ijk} = \boldsymbol{x}_{ijk}^\top \boldsymbol{\beta} + f_j(w_{ijk}) + U_i + \varepsilon_{ijk}.$$

The functions may have separate smoothing parameters or a common one. A shared value asserts that the two curves have comparable roughness, while still permitting different shapes.

The estimated socioeconomic curves in Figure 31 correspond to students outside the minority group on the left and students in the minority group on the right. They differ most in the sparsely observed low socioeconomic region. The widening uncertainty there is at least as important as the difference between the fitted means.

Smooths can also depend on several variables. A two dimensional effect $f(w_1, w_2)$ penalizes curvature across a surface, often through **thin plate** or **tensor product** bases. The mathematics example can combine individual socioeconomic status with the school's mean socioeconomic status. The infant mortality example similarly models log mortality as a smooth surface of log income

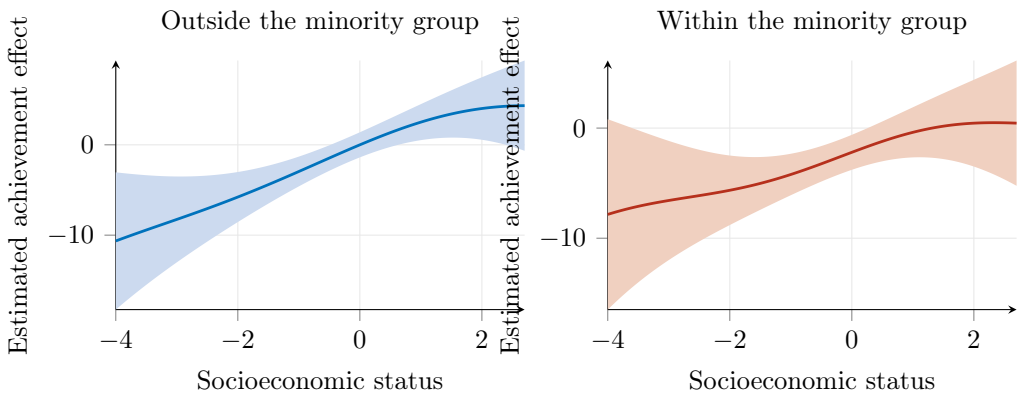

Figure 31

and income inequality. The fitted surface in Figure 32 displays a steep improvement at low income and a flatter relationship among wealthier countries, with inequality modifying the pattern.

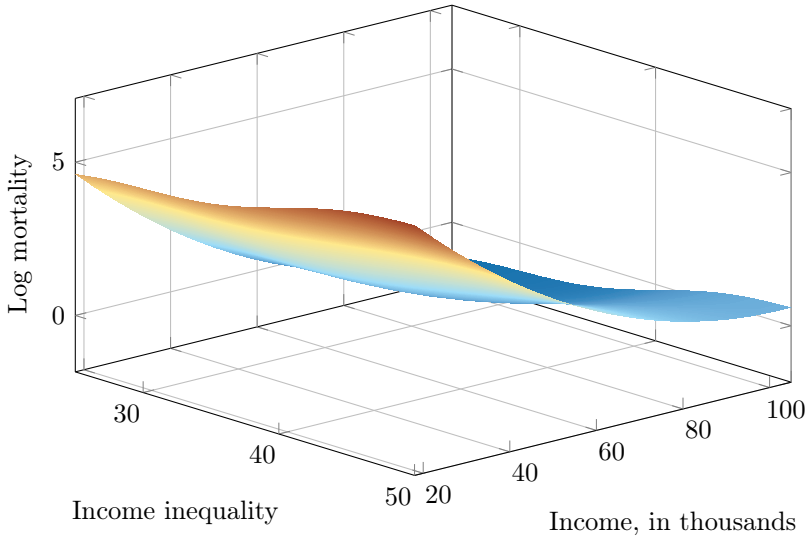

Figure 32

Smooth surfaces require data throughout the region on which they are interpreted. Predictions near the edge, or in combinations of income and inequality with little support, depend strongly on the basis and penalty. A polished contour plot should therefore be accompanied by the observed design or another indication of data density.

The spline framework is developed in detail by Eilers and Marx. A state space formulation of nonparametric smoothing is given by Brown and de Jong.

Bayesian Smoothing and Generalized Additive Mixed Models

Bayesian smoothing makes the regularization assumptions explicit. On an ordered grid, a **first order random walk prior** satisfies

$$f_{r+1} - f_r \sim \mathcal{N}(0, \sigma_f^2), \tag{8.6}$$

while a **second order random walk prior** satisfies

$$f_{r+2} - 2f_{r+1} + f_r \sim \mathcal{N}(0, \sigma_f^2). \tag{8.7}$$

The first order base model at $\sigma_f = 0$ is a constant. The second order base model is a straight line, since all second differences vanish. Thus a second order prior is usually preferable when linear regression is regarded as the scientifically reasonable default.

Both priors are intrinsic: adding a constant to a first order walk does not change its increments, and adding a straight line to a second order walk does not change its second differences. Appropriate sum to zero and trend constraints separate the smooth from the intercept and linear terms. Penalized complexity priors can be placed on σ_f to shrink toward the chosen base model.

Monthly Ontario death counts illustrate the construction. If Y_t is the number of deaths and O_t is the number of days in month t, then

$$Y_t \mid f_t, V_t \sim \text{Poisson}(O_t \lambda_t), \tag{8.8}$$

$$\log(\lambda_t) = \boldsymbol{x}_t^\top \boldsymbol{\beta} + f_t + V_t,$$

where $\boldsymbol{x}_t$ contains month indicators,

$$f_t - 2f_{t-1} + f_{t-2} \sim \mathcal{N}(0, \sigma_f^2) \qquad \text{and} \qquad V_t \sim \mathcal{N}(0, \sigma_V^2).$$

The offset $\log O_t$ is essential: without it, February appears to have fewer deaths partly because it has fewer exposure days. The independent effect V_t represents influenza outbreaks and other extra-Poisson variation not appropriately attributed to a smooth long term trend.

A representative prior calibration is

$$\mathbb{P}(\sigma_f > 0.02) = 0.05 \qquad \text{and} \qquad \mathbb{P}(\sigma_V > 0.2) = 0.05.$$

The posterior mean standard deviations are 0.00013 for the second order innovations and 0.0281 for independent monthly variation. Their 95% credible intervals are $(0.00006, 0.00024)$ and $(0.0255, 0.0308)$, respectively. Small second differences are sufficient to generate a substantial long

term trend because they accumulate across many months.

The posterior median long term relative rate and its pointwise 95% credible band are shown in Figure 33. The independent monthly component is excluded from this curve so that the slowly varying signal remains visible.

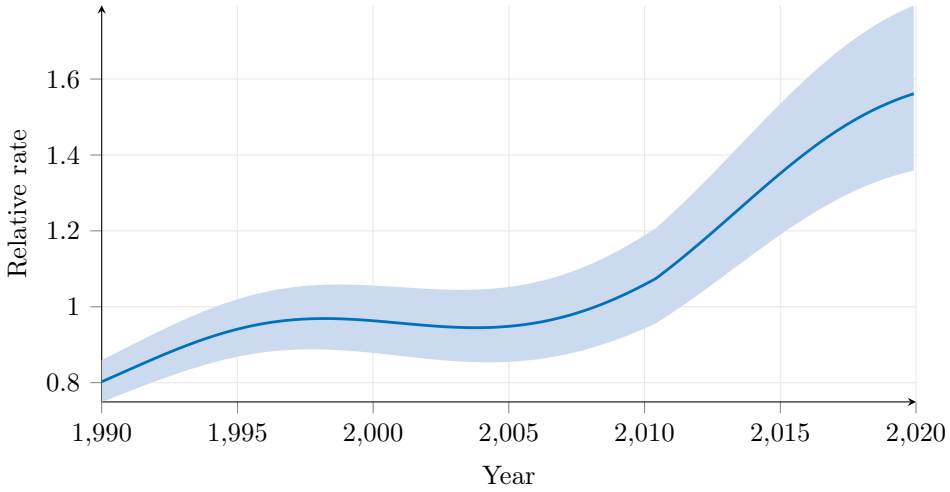

Figure 33

Eliminating V_t forces the smooth f_t to absorb every departure from the Poisson mean. The fitted innovation standard deviation then becomes much larger, the trend becomes implausibly erratic, and forecasts exaggerate long term instability. This is a sampling model failure rather than a defect in smoothing itself.

Prediction within the observed range is **interpolation**. **Forecasting** beyond the range is **extrapolation** and depends critically on the null space of the penalty. A second order random walk continues approximately linearly, while its uncertainty grows as unobserved innovations accumulate. To obtain forecasts, append future rows with missing responses, extend the latent time index, retain the month and exposure variables, and compute posterior predictive summaries.

The resulting monthly forecasts continue the estimated long term trajectory and seasonal pattern. The vertical boundary marks the end of the observed series; uncertainty widens beyond it.

The forecast distribution and its widening uncertainty appear in Figure 34.

Random walk smooths also permit functions specific to each group. For the mathematics data,

$$Y_{ijk} = \boldsymbol{x}_{ijk}^{\top}\boldsymbol{\beta} + U_i + f_j(w_{ijk}) + \varepsilon_{ijk} \qquad \text{and} \qquad U_i \sim \mathcal{N}(0, \sigma_U^2),$$

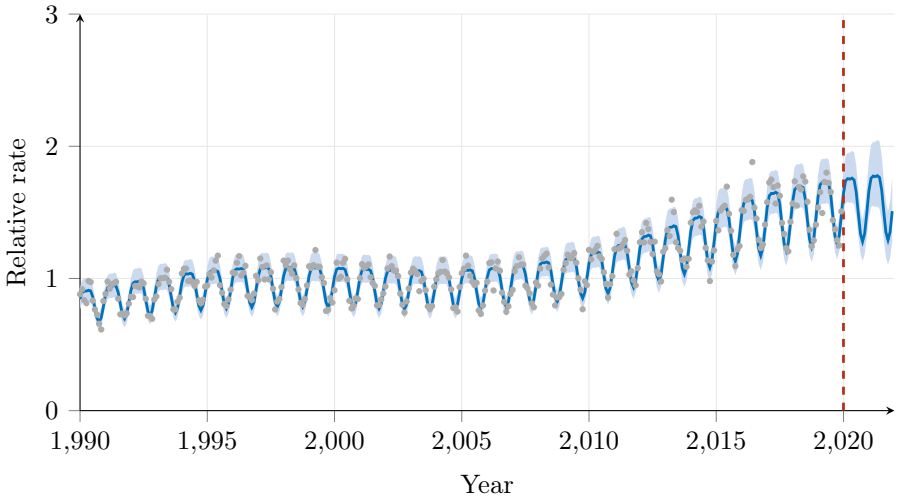

Figure 34

where j indexes minority status. Replicated first or second order random walks allow f_j to differ between groups while sharing a common innovation variance. A first order fit estimates smooth standard deviation about 0.406; the corresponding second order value is about 0.0092. These numbers are not directly comparable because first and second differences have different units and null spaces.

A **generalized additive mixed model** combines smooth functions with structured random effects:

$$Y_{ij} \sim \mathcal{F}(\mu_{ij}, \boldsymbol{\theta}), \qquad g(\mu_{ij}) = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + f(w_{ij}) + U_i, \qquad \text{and} \qquad \boldsymbol{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{8.9}$$

The smooth and the cluster effect solve different problems. The former represents a nonlinear mean relationship; the latter represents dependence among observations from the same group. Omitting either can distort uncertainty in the other.

Penalized likelihood and fully Bayesian approaches share the same model components but answer uncertainty questions differently. A penalized fit estimates smoothing parameters by generalized cross validation or restricted likelihood and then commonly conditions on those estimates. A Bayesian fit places priors on smoothing and variance parameters and integrates over their posterior uncertainty. The Bayesian calculation is more explicit and extends naturally to censored responses, spatial effects, and complex dependence, but it demands careful prior calibration and can fail when the model contains more complexity than the data support.

The label **nonparametric** should not be mistaken for freedom from assumptions. Spline bases

may contain many anonymous coefficients, yet the basis dimension, penalty, boundary behaviour, smoothing selection, link, response distribution, and dependence structure are all substantive assumptions. Scientific knowledge should determine which relationships deserve flexible functions before those functions are inspected in the data.

Lecture 9 — Wednesday, March 13, 2019

5. Spatial Models and Retrospective Sampling

Gaussian Random Fields and Covariance Models

Geostatistics concerns an underlying spatial surface observed only at a finite set of locations. The principal objectives are to model spatial dependence, estimate its parameters, predict the surface between observations, and quantify uncertainty in those predictions.

Soil mercury illustrates the setting. Concentration is measured at sampling locations, while elevation, nighttime light, vegetation, and land cover are available more broadly. The scientific target is not merely the observed sample; it is the latent mercury surface throughout the region. The sampling locations and the spatially varying land cover information are shown in Figure 35.

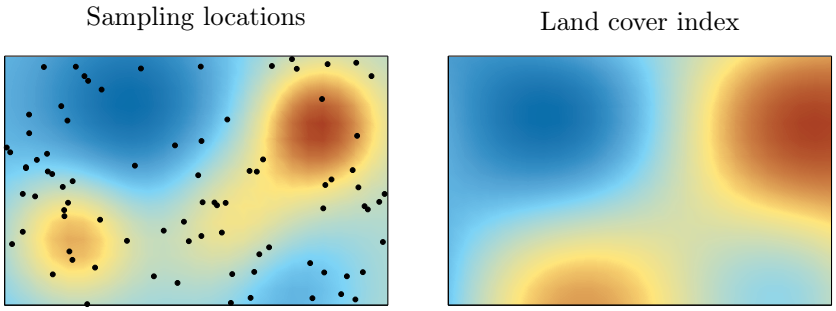

Figure 35

Different data generating mechanisms lead to different spatial models. Continuous mercury measurements motivate Gaussian geostatistics. Radiation particle counts on Rongelap Island require a non-Gaussian response model, as developed by Diggle, Moyeed, and Tawn. Locations of events such as homicides are themselves random and call for a **spatial point process**. These extensions share a latent spatial surface but use different observation models.

Definition 9.1 A spatial process $\{U(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is a **Gaussian random field** if, for every finite collection $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$,

$$\begin{pmatrix} U(\boldsymbol{s}_1) \\ \vdots \\ U(\boldsymbol{s}_N) \end{pmatrix}$$

has a multivariate normal distribution.

A Gaussian random field is specified completely by its **mean function**

$$m(\boldsymbol{s}) = \mathbb{E}[U(\boldsymbol{s})]$$

and **covariance function**

$$C(\boldsymbol{s}, \boldsymbol{t}) = \mathrm{Cov}(U(\boldsymbol{s}), U(\boldsymbol{t})).$$

The field is **second order stationary** when its mean is constant and its covariance depends only on the displacement $\boldsymbol{h} = \boldsymbol{s} - \boldsymbol{t}$. It is **isotropic** when the covariance depends only on distance $\|\boldsymbol{h}\|$. Stationarity is an assumption about invariance under translation; isotropy adds invariance under rotation.

A covariance model must produce a **positive semidefinite matrix** for every finite set of locations. Not every decreasing function of distance has this property. The exponential model

$$C(\boldsymbol{h}) = \sigma^2 \exp\left(-\frac{\|\boldsymbol{h}\|}{\phi}\right)$$

is valid and produces continuous but rough surfaces. Larger ϕ gives more persistent correlation and smoother large scale appearance.

The **Matérn family** separates correlation range from local smoothness:

$$\rho(\boldsymbol{h}; \phi, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{8\nu}\|\boldsymbol{h}\|}{\phi}\right)^{\nu} K_\nu \left(\frac{\sqrt{8\nu}\|\boldsymbol{h}\|}{\phi}\right), \tag{9.1}$$

where K_ν is the modified Bessel function of the second kind, $\phi > 0$ is a range parameter, and $\nu > 0$ controls smoothness. This scaling makes ϕ a comparable practical range across values of ν.

When $\nu = 1/2$, (9.1) reduces to an exponential correlation after the corresponding range rescaling. As ν grows, the field becomes smoother; a Matérn field is differentiable m times in mean square precisely when $\nu > m$. Range and smoothness are distinct: range controls the distance over which values remain associated, while ν controls local differentiability.

The panels in Figure 36 compare correlations and realizations as range and smoothness change. Fields with short range vary rapidly across space, while fields with large ν have smoother local texture.

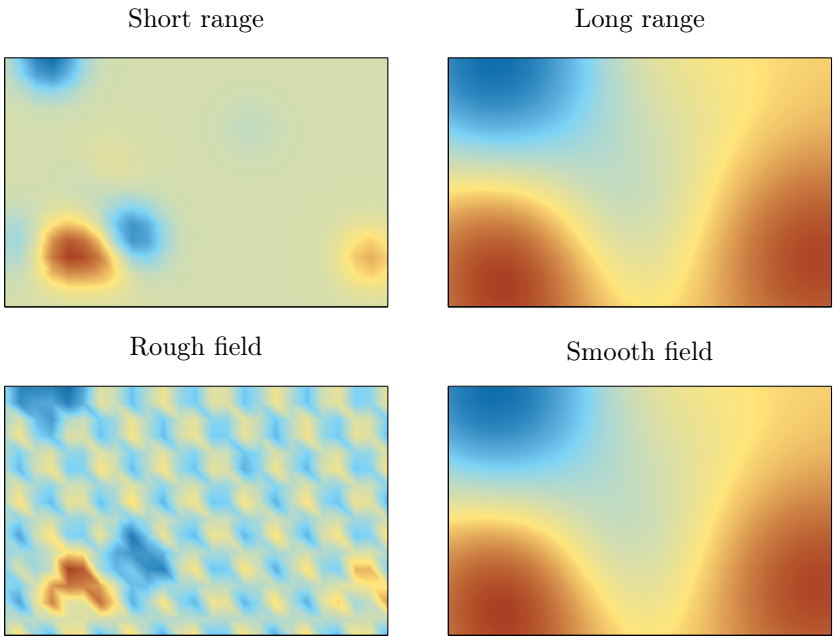

Figure 36

Isotropy may fail when transport has a preferred direction, as with prevailing wind, water flow, or geological structure. **Geometric anisotropy** transforms the displacement before applying an isotropic correlation. Let

$$\mathbf{R}(\psi) = \begin{pmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{pmatrix} \qquad \text{and} \qquad \mathbf{D} = \begin{pmatrix} 1/\phi_1 & 0 \\ 0 & 1/\phi_2 \end{pmatrix}.$$

The **anisotropic distance** is

$$d_A(\boldsymbol{h}) = \|\mathbf{D}\mathbf{R}(\psi)\boldsymbol{h}\|, \tag{9.2}$$

and the correlation is $\rho_0\{d_A(\boldsymbol{h})\}$ for a valid isotropic correlation ρ_0. Equal correlation contours are ellipses with principal ranges ϕ_1 and ϕ_2. Under the displayed convention, their principal axes are obtained by rotating the coordinate axes through $-\psi$.

The directional distortion created by anisotropy is illustrated in Figure 37.

Anisotropy introduces a direction and a range ratio in addition to an overall range. These parameters should be estimated only when the locations span enough directions and distances. Apparent

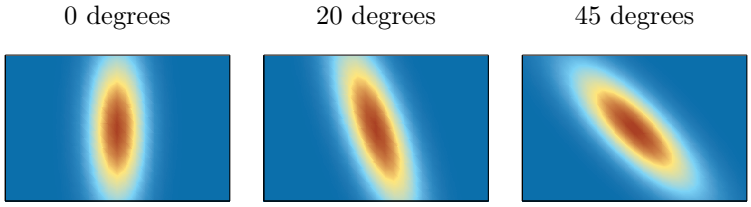

Figure 37

directional structure may instead be a mean trend that belongs among the explanatory variables.

Geostatistical Inference and Kriging

A **linear Gaussian geostatistical model** separates broad scale covariate effects, residual spatial variation, and independent noise:

$$Y_i = \mu + \boldsymbol{x}(\boldsymbol{s}_i)^\top \boldsymbol{\beta} + U(\boldsymbol{s}_i) + \varepsilon_i \qquad \text{and} \qquad \varepsilon_i \sim \mathcal{N}(0, \tau^2), \tag{9.3}$$

where

$$\text{Cov}(U(\boldsymbol{s}+\boldsymbol{h}), U(\boldsymbol{s})) = \sigma^2 \rho(\boldsymbol{h}; \phi, \nu).$$

The nugget variance τ^2 represents measurement error or variation at a spatial scale finer than the sampling design. The spatial variance σ^2 measures residual surface variation after the covariates have been included.

Let

$$\boldsymbol{Y} = (Y_1, \ldots, Y_N)^\top \qquad \text{and} \qquad \boldsymbol{U} = \{U(\boldsymbol{s}_1), \ldots, U(\boldsymbol{s}_N)\}^\top.$$

Then

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \qquad \text{and} \qquad \mathbf{V} = \sigma^2 \mathbf{R}(\phi, \nu) + \tau^2 \mathbf{I}_N. \tag{9.4}$$

The log likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}\left[N \log(2\pi) + \log|\mathbf{V}| + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})\right], \tag{9.5}$$

where $\boldsymbol{\theta}$ contains the covariance parameters.

If $\mathbf{V}$ is positive definite and $\mathbf{X}$ has full column rank, then, for fixed $\boldsymbol{\theta}$, generalized least squares gives

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{V}^{-1} \boldsymbol{y}.$$

Substitution into (9.5) profiles out $\boldsymbol{\beta}$, leaving numerical optimization over the covariance parame-

ters. Each evaluation requires a Cholesky factorization of the dense matrix $\mathbf{V}$, so direct likelihood costs order N^3 operations and order N^2 storage.

For the European soil mercury analysis, the transformed response includes elevation, nighttime light, vegetation, and land cover. With Matérn smoothness fixed at 1.5, the principal estimates are

parameter	estimate	2.5%	97.5%
elevation, per 1000	0.32	0.00	0.64
nighttime light, per 200	0.37	0.04	0.70
vegetation index	2.42	1.54	3.30
nugget standard deviation	1.43	1.16	1.76
spatial standard deviation	1.38	1.02	1.86
range, kilometres	531	344	817
Box–Cox parameter	−0.28	−0.34	−0.22

The nugget and spatial standard deviations are comparable, so neither independent noise nor residual spatial structure can be ignored. Several forest and sparse vegetation categories have lower transformed mercury than rainfed cropland.

Spatial prediction follows from conditioning a multivariate normal distribution.

Proposition 9.2 Suppose

$$\begin{pmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_A & \mathbf{C} \\ \mathbf{C}^\top & \boldsymbol{\Sigma}_B \end{pmatrix} \right).$$

Assume that $\boldsymbol{\Sigma}_B$ is positive definite. Then

$$\boldsymbol{A} \mid \boldsymbol{B} \sim \mathcal{N}\left(\boldsymbol{\mu}_A + \mathbf{C}\boldsymbol{\Sigma}_B^{-1}(\boldsymbol{B} - \boldsymbol{\mu}_B), \boldsymbol{\Sigma}_A - \mathbf{C}\boldsymbol{\Sigma}_B^{-1}\mathbf{C}^\top \right).$$

Let $\boldsymbol{g}_1, \dots, \boldsymbol{g}_M$ be prediction grid locations and let

$$\boldsymbol{U}_G = \{U(\boldsymbol{g}_1), \dots, U(\boldsymbol{g}_M)\}^\top.$$

Define $\mathbf{C}_{GY}$ by

$$\{\mathbf{C}_{GY}\}_{ki} = \sigma^2 \rho(\boldsymbol{g}_k - \boldsymbol{s}_i).$$

Applying Proposition 9.2 gives the **simple kriging equations**

$$\mathbb{E}[\boldsymbol{U}_G \mid \boldsymbol{Y} = \boldsymbol{y}] = \mathbf{C}_{GY}\mathbf{V}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \tag{9.6}$$

and

$$\operatorname{Var}(\boldsymbol{U}_G \mid \boldsymbol{Y}) = \boldsymbol{\Sigma}_G - \mathbf{C}_{GY}\mathbf{V}^{-1}\mathbf{C}_{GY}^{\top}. \tag{9.7}$$

Prediction of the latent mean surface adds $\mathbf{X}_G\boldsymbol{\beta}$ to (9.6). Prediction of a future noisy measurement additionally adds τ^2 to the appropriate diagonal variance. These are different targets and should be labelled distinctly.

The left map in Figure 38 is the predicted residual spatial effect, while the right map adds the estimated covariate contribution to obtain the predicted transformed mercury surface. Sampling locations exert more influence nearby, with the covariance model controlling how that influence decays.

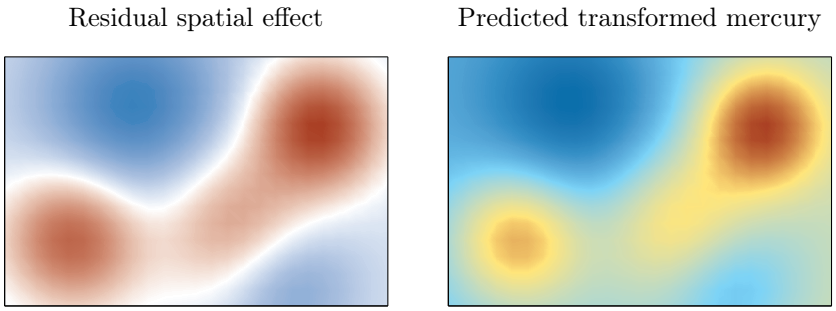

Figure 38

For a Gaussian predictive target $Z(\boldsymbol{g})$, whether a latent mean or a future response, an **exceedance probability** is

$$\mathbb{P}(Z(\boldsymbol{g}) > c \mid \boldsymbol{Y}) = 1 - \Phi\left[\frac{c - \mathbb{E}[Z(\boldsymbol{g}) \mid \boldsymbol{Y}]}{\sqrt{\operatorname{Var}(Z(\boldsymbol{g}) \mid \boldsymbol{Y})}}\right].$$

Mapping this quantity answers a threshold question more directly than mapping the posterior mean. A residual exceedance map, such as $\mathbb{P}(U(\boldsymbol{g}) > 0 \mid \boldsymbol{Y})$, instead identifies locations unusually high relative to the covariates.

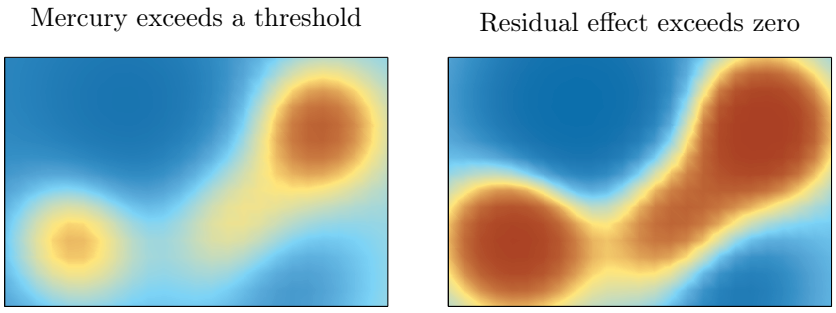

Figure 39

The corresponding exceedance probabilities are mapped in Figure 39.

The equations that plug in covariance estimates treat them as known and therefore omit parameter uncertainty. Restricted likelihood, parametric bootstrap, or Bayesian integration can restore some of that uncertainty. Direct dense matrix methods work for several thousand observations; larger data sets require sparse representations, low rank approximations, covariance tapering, or other scalable constructions.

The model based treatment of Gaussian geostatistics and prediction is developed by Diggle and Ribeiro.

Lecture 10 — Wednesday, March 20, 2019

Generalized Linear Geostatistical Models

Gaussian geostatistics can be extended to counts, proportions, and binary responses by combining a generalized linear model with a spatially correlated latent field. Let Y_i be observed at location $\boldsymbol{s}_i$, and suppose that the observations are conditionally independent given a Gaussian random field U. A **generalized linear geostatistical model** has the form

$$Y_i \mid U(\boldsymbol{s}_i), \boldsymbol{\beta}, \vartheta \sim \mathcal{F}(\mu_i, \vartheta), \tag{10.1}$$

$$g(\mu_i) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + U(\boldsymbol{s}_i), \tag{10.2}$$

where g is an appropriate link function, ϑ denotes any response dispersion parameter, and U has its own covariance parameters in a model such as the Matérn family from Section 9. The conditional independence in (10.1) does not imply marginal independence: after U is integrated out, nearby responses remain associated through the shared field.

The latent field has three distinct roles. It represents omitted spatially structured covariates, accounts for residual dependence, and permits prediction at unobserved locations. The interpretation of the regression coefficients is conditional on that latent field. For example, in a binary model,

$$Y_i \mid U(\boldsymbol{s}_i) \sim \text{Bernoulli}(p_i) \qquad \text{and} \qquad \text{logit}(p_i) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + U(\boldsymbol{s}_i),$$

so $\exp(\beta_j)$ is a conditional odds ratio for a one unit increase in x_{ij}, holding the other covariates and the residual spatial effect fixed.

The following Toronto example records whether each inspected coffee shop offered a latte. The points are not uniformly distributed over the city, and neighbouring establishments may resemble one another because of unmeasured local characteristics.

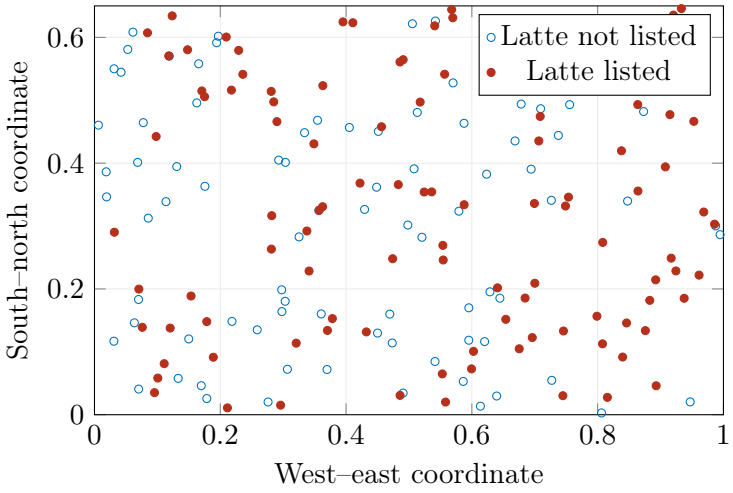

Figure 40

The locations of the sampled coffee shops are shown in Figure 40.

Let $Y(\boldsymbol{s}) = 1$ if the shop at $\boldsymbol{s}$ offers a latte. One fitted model was

$$\operatorname{logit}(p(\boldsymbol{s})) = \beta_0 + \beta_1 \log(\text{income}(\boldsymbol{s})) + \beta_2 \,\text{ford}(\boldsymbol{s}) + U(\boldsymbol{s}),$$

where $\text{ford}(\boldsymbol{s})$ indicates one side of the city and U is a Matérn field. The posterior summaries were as follows.

Quantity	Posterior mean	95% credible interval
$\exp(\beta_0)$	0.32	$(0.26, 0.36)$
$\exp(\beta_1)$	2.46	$(1.79, 3.35)$
$\exp(\beta_2)$	2.83	$(2.13, 3.87)$
Spatial range, in kilometres	6.36	$(2.01, 17.18)$
Spatial standard deviation	0.01	$(0.00, 0.03)$

The covariate associations are substantial on the odds scale, whereas the residual spatial standard deviation is close to zero after adjustment. The estimated range is consequently weakly identified: a field with almost no residual amplitude cannot provide much information about the distance over which that field remains correlated. This is a useful warning against interpreting a covariance range without first examining the associated variance.

The upper left panel of Figure 41 maps the fitted prevalence $\mathbb{E}[p(\boldsymbol{s}) \mid \boldsymbol{Y}]$; the upper right panel maps the multiplicative residual surface $\mathbb{E}[\exp(U(\boldsymbol{s})) \mid \boldsymbol{Y}]$. The first combines covariate and residual effects, while the second isolates unexplained spatial structure.

For a prevalence threshold c, the posterior exceedance probability

$$\mathbb{P}(p(\boldsymbol{s}) > c \mid \boldsymbol{Y})$$

identifies places where a scientifically meaningful threshold is likely to be exceeded. A corresponding probability for $U(\boldsymbol{s})$ asks whether the unexplained spatial effect is unusually large. These are different questions and need not produce similar maps.

The lower left panel maps the prevalence exceedance probability at $c = 0.6$, while the lower right panel maps the probability that the multiplicative residual exceeds one.

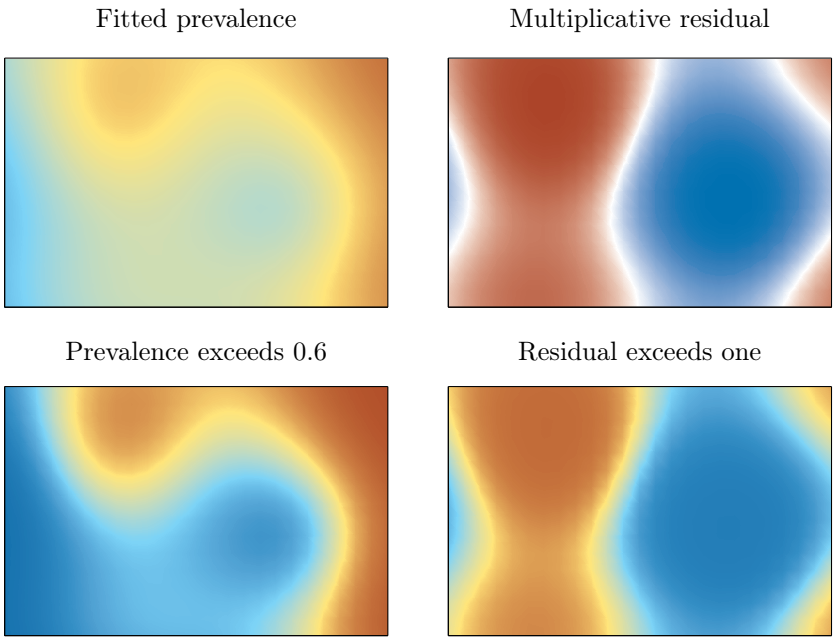

Figure 41

Inference is more difficult than in the Gaussian model because the latent field cannot generally be integrated out in closed form. Markov chain Monte Carlo, Laplace approximation, or integrated nested Laplace approximation can be used. The likelihood contribution conditional on $\boldsymbol{U}$ remains simple,

$$p(\boldsymbol{y} \mid \boldsymbol{U}, \boldsymbol{\beta}, \vartheta) = \prod_{i=1}^{n} p(y_i \mid U(\boldsymbol{s}_i), \boldsymbol{\beta}, \vartheta),$$

but the Gaussian prior for $\boldsymbol{U}$ introduces a dense covariance matrix in an ordinary geostatistical representation. Sparse precision models provide a computationally preferable alternative.

Lattice Models and Gaussian Markov Random Fields

Geostatistical models index a field by every point in a continuous region. A **lattice model** instead associates one random variable U_i with each cell or region i. The lattice may be a regular grid, or it may be an irregular collection of administrative areas. Its essential ingredient is a neighbourhood graph: write $i \sim j$ when regions i and j are neighbours, and let w_{ij} record the corresponding edge.

A **Markov random field** satisfies the local Markov property

$$U_i \perp\!\!\!\perp \{U_j : j \notin (\partial i \cup \{i\})\} \mid \{U_j : j \in \partial i\},$$

where ∂i is the set of neighbours of region i. Thus, conditional on its neighbours, a cell is independent of all more distant cells. The neighbourhood need not be based on a shared border; distance, transportation links, or scientific connectivity may define a more appropriate graph.

The most common Gaussian construction is the **intrinsic conditional autoregressive model**. If $n_i = \sum_j w_{ij}$ and

$$\overline{U}_i = \frac{1}{n_i} \sum_{j=1}^{n} w_{ij} U_j,$$

then its full conditional distributions are

$$U_i \mid \boldsymbol{U}_{-i}, \tau \sim \mathcal{N}\left(\overline{U}_i, \frac{1}{\tau n_i}\right). \tag{10.3}$$

The conditional mean smooths U_i toward the average of its neighbours. A region with many neighbours has smaller conditional variance because its local mean is supported by more adjacent values.

Full conditional distributions cannot be chosen independently; they must be compatible with a joint distribution. For example, two conditionals that force $U_1 < U_2$ and $U_2 < U_1$, respectively, cannot both hold. For a connected undirected graph with symmetric weights, the conditionals in (10.3) correspond to the unnormalized density

$$p(\boldsymbol{u} \mid \tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2} \sum_{i<j} w_{ij}(u_i - u_j)^2\right). \tag{10.4}$$

Only neighbouring differences appear in (10.4). Adding the same constant to every u_i leaves the density unchanged, so its integral over $\mathbb{R}^n$ is infinite. The intrinsic conditional autoregressive model is therefore improper. A constraint such as

$$\sum_{i=1}^{n} U_i = 0$$

separates the overall intercept from the spatial effect and makes the model identifiable. A graph with several connected components requires one constraint for each component.

Let $\mathbf{W} = (w_{ij})$ and $\mathbf{D} = \operatorname{diag}(n_1, \ldots, n_n)$. The intrinsic precision matrix is

$$\mathbf{Q}_{\text{ICAR}} = \tau(\mathbf{D} - \mathbf{W}).$$

This matrix is sparse because $Q_{ij} = 0$ whenever i and j are not neighbours. It is singular because $(\mathbf{D} - \mathbf{W})\mathbf{1} = \mathbf{0}$, which is the matrix expression of invariance under constant shifts.

Theorem 10.1 Suppose that $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$, where $\mathbf{Q}$ is positive definite. For distinct indices i and j, Y_i and Y_j are conditionally independent given all remaining coordinates if and only if $Q_{ij} = 0$.

Proof. Completing the square in the Gaussian density gives

$$Y_i \mid \boldsymbol{Y}_{-i} = \boldsymbol{y}_{-i} \sim \mathcal{N}\left(-\frac{1}{Q_{ii}} \sum_{k \neq i} Q_{ik} y_k, \frac{1}{Q_{ii}} \right). \tag{10.5}$$

Forward implication. Suppose that Y_i and Y_j are conditionally independent given the remaining coordinates. The conditional distribution in (10.5) cannot depend on y_j. Its coefficient is $-Q_{ij}/Q_{ii}$, so $Q_{ij} = 0$.

Reverse implication. Suppose that $Q_{ij} = 0$. The mean and variance in (10.5) then do not depend on y_j. Hence the conditional distribution of Y_i, given Y_j and all remaining coordinates, equals its conditional distribution given only the remaining coordinates. This is precisely the required conditional independence. □

Theorem 10.1 explains why a Gaussian Markov random field is represented through a precision matrix rather than a covariance matrix. Local dependence becomes sparsity, and sparse Cholesky

factorization can be far less expensive than factorizing the dense covariance matrices of ordinary geostatistics.

A **proper conditional autoregressive model** discounts the neighbour average:

$$U_i \mid \boldsymbol{U}_{-i}, \tau, \varphi \sim \mathcal{N}\left(\frac{\varphi}{n_i}\sum_j w_{ij}U_j, \frac{1}{\tau n_i}\right) \qquad \text{and} \qquad \mathbf{Q} = \tau(\mathbf{D} - \varphi\mathbf{W}). \tag{10.6}$$

For a binary adjacency matrix on a connected graph, $0 \leq \varphi < 1$ gives a positive definite precision matrix. More generally, the admissible interval is determined by the eigenvalues of $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. The intrinsic model is recovered at the singular boundary $\varphi = 1$.

The first panel of Figure 42 illustrates an intrinsic field on a regular grid; its absolute level is arbitrary, but neighbouring differences are smooth. The remaining panels illustrate how the dependence parameter and precision affect proper fields. Stronger dependence produces greater local persistence, while greater precision reduces marginal variation. All four panels use the same colour scale so that this reduction in variation remains visible.

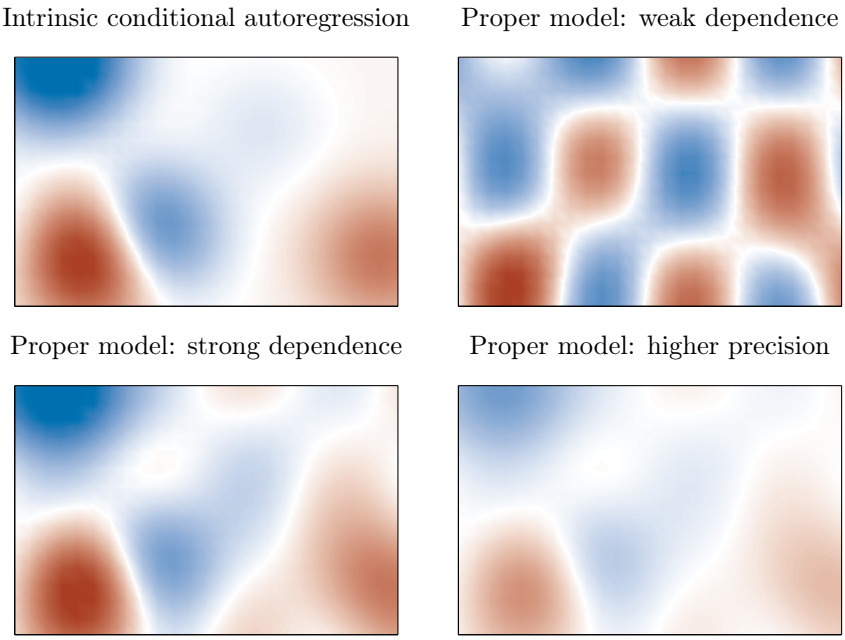

Figure 42

These fields are closely connected to Matérn models. On a fine regular grid, the first order proper conditional autoregressive model approximates a Matérn field with smoothness near zero. More

generally, the stochastic partial differential equation

$$(\kappa^2 - \Delta)^{\alpha/2}\{\tau U(\boldsymbol{s})\} = \mathcal{W}(\boldsymbol{s}) \tag{10.7}$$

has a Matérn solution, where $\mathcal{W}$ is Gaussian white noise, Δ is the Laplacian, and $\alpha > d/2$, so the smoothness in d dimensions is $\nu = \alpha - d/2 > 0$. A finite element discretization of (10.7) produces a Gaussian Markov random field with sparse precision. This connection preserves continuous space interpretation while supplying the computational advantages of a sparse lattice representation.

Higher order neighbourhoods arise from higher powers of a discrete differential operator. On an interior cell of a square lattice, for example, a second order stencil can involve the four nearest neighbours, the four diagonal neighbours, and the four cells two steps away. Boundary rows must be modified because fewer neighbours are available. The resulting precision remains sparse even though its neighbourhood is larger.

The foundational conditional autoregressive construction is due to Besag; the sparse Gaussian theory is developed by Rue and Held; and the stochastic partial differential equation representation is due to Lindgren, Rue, and Lindström.

Spatial Point Processes and Areal Models

A **spatial point pattern** records the event locations themselves. If $N(A)$ is the number of events in a bounded region A, a **Poisson point process** with nonnegative, locally integrable intensity $\Lambda(\boldsymbol{s})$ satisfies

$$N(A) \sim \text{Poisson}\left(\int_A \Lambda(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}\right),$$

and counts in disjoint regions are independent. The intensity is a rate per unit area, not a probability. In particular, $\Lambda(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}$ approximates the expected number of events in a small area around $\boldsymbol{s}$.

When the population at risk is not spatially uniform, write

$$\Lambda(\boldsymbol{s}) = O(\boldsymbol{s})\lambda(\boldsymbol{s}),$$

where $O(\boldsymbol{s})$ is a known exposure surface and $\lambda(\boldsymbol{s})$ is the event rate per unit exposure. A **log-Gaussian Cox process** makes the log rate a latent Gaussian field:

$$\log(\lambda(\boldsymbol{s})) = \mu + \boldsymbol{x}(\boldsymbol{s})^{\mathsf{T}}\boldsymbol{\beta} + U(\boldsymbol{s}). \tag{10.8}$$

Conditional on U, the process is Poisson. Marginally, the random intensity induces clustering

beyond that explained by the measured covariates and exposure.

For computation, partition the study region into small cells A_k. If the covariates and latent field are nearly constant within a cell, then

$$N(A_k) \mid U \;\dot{\sim}\; \text{Poisson}\left[|A_k| O(\boldsymbol{s}_k) \exp\Big(\mu + \boldsymbol{x}(\boldsymbol{s}_k)^{\mathsf{T}}\boldsymbol{\beta} + U(\boldsymbol{s}_k)\Big)\right].$$

This approximation turns the point process likelihood into a large Poisson regression. A sparse Gaussian Markov random field representation of U then makes integrated nested Laplace approximation feasible.

As an illustration, consider the locations of murders in Toronto over thirty years. A fitted model used log income, log population density, and log nighttime light intensity as covariates. The observed events and a local enlargement appear in Figure 43.

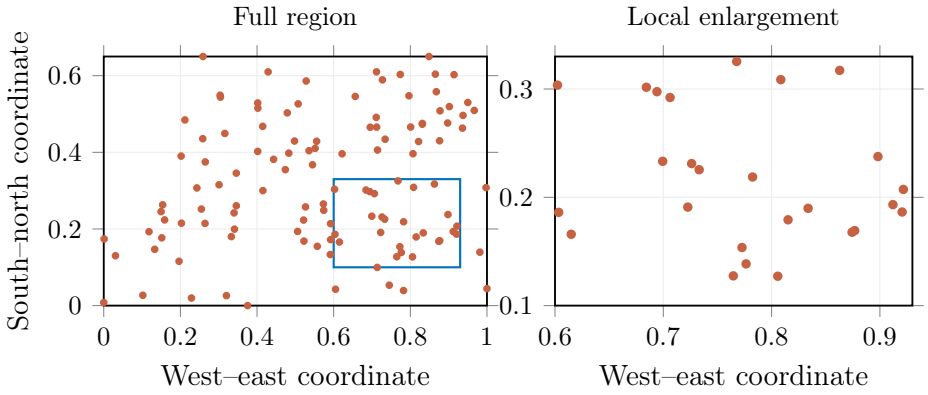

Figure 43

The posterior summaries were

Quantity	Posterior mean	95% credible interval
Intercept	-1.92	$(-6.42, 2.59)$
Log income	-1.53	$(-1.91, -1.15)$
Log population density	0.75	$(0.62, 0.88)$
Log nighttime light intensity	0.55	$(0.32, 0.77)$
Spatial range, in metres	828.68	$(691.86, 989.66)$
Spatial standard deviation	0.99	$(0.90, 1.08)$

Because income, population density, and light intensity are on logarithmic scales, their coefficients are elasticities of the conditional event rate. For example, doubling income multiplies the fitted

rate by $2^{-1.53}$, approximately 0.35, after the other variables and latent field are held fixed. This is an adjusted spatial association, not a causal effect. The sizeable residual standard deviation shows that substantial local variation remains after adjustment.

The posterior mean rate surface is shown first, followed by the residual multiplicative surface $\mathbb{E}[\exp(U(\boldsymbol{s})) \mid \boldsymbol{Y}]$. The broad rate pattern reflects the measured predictors and latent field, whereas the residual surface reveals departures not explained by the predictors. The point intensity would additionally multiply the fitted rate by the exposure surface.

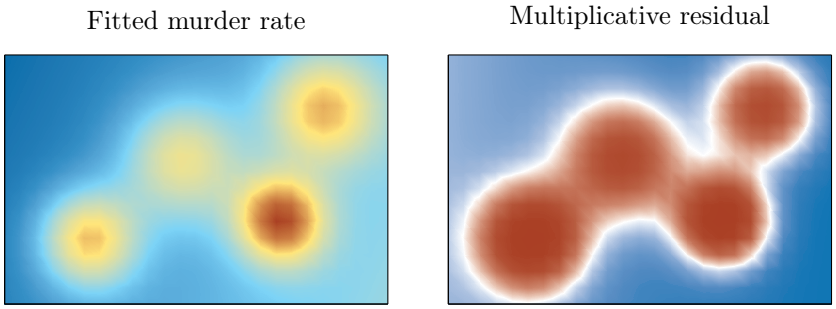

Figure 44

The fitted murder rate surface and residual multiplicative surface are compared in Figure 44.

The same framework can compare spatial patterns specific to each disease. In a joint analysis of lung and thyroid cancer, higher income and larger Asian population proportion were negatively associated with lung cancer rate after adjustment, while the Asian population proportion was positively associated with thyroid cancer rate. Several other intervals included zero. The fitted range was longer for lung cancer, but the intervals were wide. Separate latent fields allow the two cancers to exhibit different residual spatial scales.

The disease-specific spatial surfaces are displayed in Figure 45.

When exact event locations are unavailable, the data may consist only of counts Y_i for regions A_i. If

$$E_i = \int_{A_i} O(\boldsymbol{s}) \, \mathrm{d}\boldsymbol{s}$$

is the expected count under a reference rate, an **areal Poisson model** is

$$Y_i \mid \lambda_i \sim \mathrm{Poisson}(E_i \lambda_i), \tag{10.9}$$

where λ_i is a relative risk. The **crude standardized incidence ratio** Y_i/E_i is unstable when E_i is small. Hierarchical spatial smoothing borrows strength among neighbouring regions while

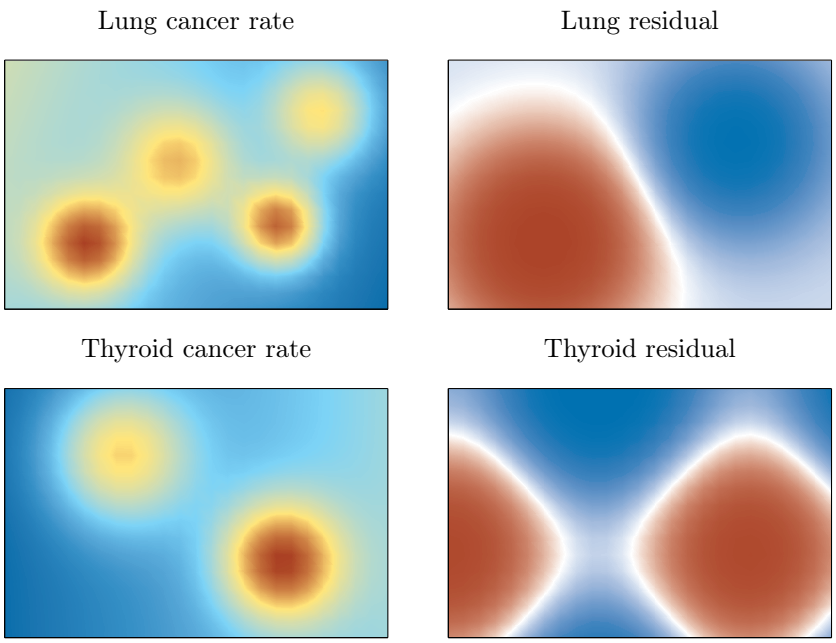

Figure 45

retaining the expected count as an offset.

The Kentucky larynx cancer data illustrate this aggregation. The observed and expected counts vary with both disease risk and population size, so the two maps must be interpreted together.

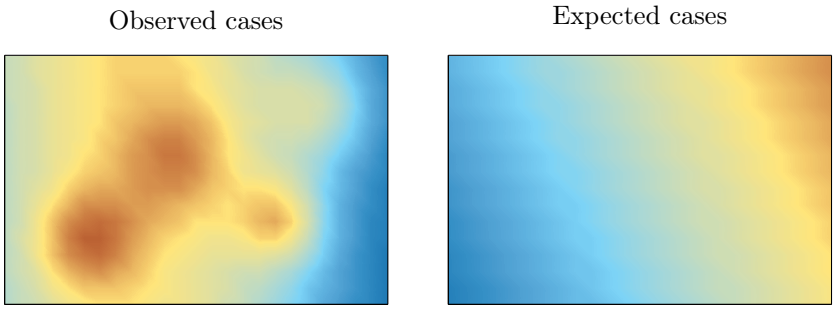

Figure 46

Observed and expected Kentucky county counts are mapped together in Figure 46.

The **Besag–York–Mollié (BYM) model** combines structured and unstructured residual varia-

tion:

$$Y_i \mid \lambda_i \sim \text{Poisson}(E_i \lambda_i), \tag{10.10}$$

$$\log(\lambda_i) = \mu + \boldsymbol{x}_i^{\mathsf{T}} \boldsymbol{\beta} + W_i + V_i, \tag{10.11}$$

where $\boldsymbol{W}$ follows an intrinsic conditional autoregressive model and the V_i are independent normal random variables. The two components serve different purposes. The structured term smooths over the adjacency graph; the independent term allows a single region to depart from its neighbours.

The original parametrization assigns separate variance parameters to W_i and V_i. Their marginal scales are difficult to compare because the intrinsic component depends on the graph. A scaled reparametrization, often called the **scaled Besag–York–Mollié model**, writes

$$b_i = \sigma_b \left\{ \sqrt{1-\omega}\, v_i + \sqrt{\omega}\, u_i^{\star} \right\}, \tag{10.12}$$

where $v_i \sim \mathcal{N}(0,1)$, $\boldsymbol{u}^{\star}$ is a scaled intrinsic conditional autoregressive field, σ_b is the total residual standard deviation, and $0 \le \omega \le 1$ is the proportion of residual variance assigned to the spatially structured component. This parametrization separates overall variability from spatial allocation and supports interpretable penalized complexity priors.

For the Kentucky data, a model including county poverty percentage produced a posterior mean coefficient near 0.01, with a 95% credible interval extending from approximately -0.01 to 0.03. The estimated total residual standard deviation was 0.10, with interval $(0.01, 0.37)$, and the estimated spatial proportion was 0.35, with interval $(0.00, 0.92)$. The wide interval for ω shows that these data do not clearly distinguish structured from independent heterogeneity.

The fitted relative risks and residual effects in Figure 47 are smoother than the crude ratios because counties share information through the hierarchy. Smoothing should not be confused with erasing genuine local variation: an isolated extreme count can still be represented by the independent component.

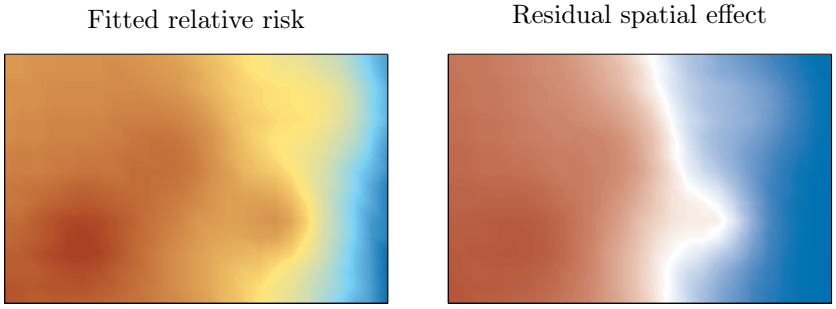

Figure 47

Posterior exceedance probabilities communicate whether a threshold is credibly crossed. For example,

$$\mathbb{P}(\lambda_i > 1.1 \mid \boldsymbol{Y}) \qquad \text{and} \qquad \mathbb{P}(b_i > \log(1.01) \mid \boldsymbol{Y})$$

address elevated relative risk and elevated residual effect, respectively. The maps in Figure 48 demonstrate that a region can have high fitted risk because of measured covariates without having an unusually large residual effect.

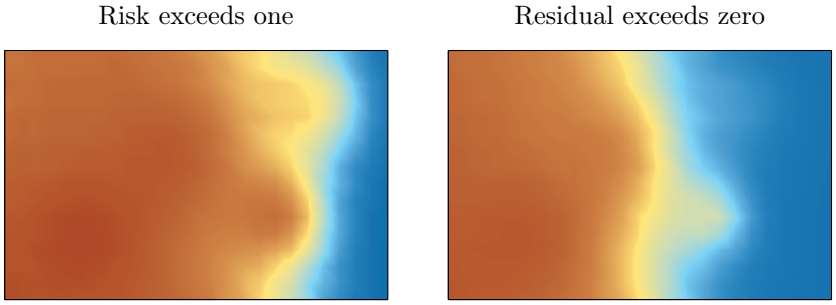

Figure 48

Areal outcomes need not be Poisson. If Y_i of m_i individuals in region i have a binary outcome, then

$$Y_i \mid p_i \sim \mathrm{Bin}(m_i, p_i) \qquad \text{and} \qquad \mathrm{logit}(p_i) = \mu + \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + b_i,$$

with b_i given by (10.12). A county level analysis of votes in the 2016 Georgia primary used demographic and socioeconomic covariates together with a scaled spatial effect. After adjustment, the posterior mean coefficient of the proportion male was 1.399, with interval $(0.479, 2.326)$, and the coefficient of log income was -0.744, with interval $(-0.927, -0.562)$. A ten percentage point increase in the proportion male therefore multiplies the fitted odds by $\exp(0.1 \times 1.399)$, approximately 1.15, while doubling median income multiplies them by $2^{-0.744}$, approximately 0.60. These are adjusted associations at the county level; they do not describe individual voting behaviour and have no causal interpretation.

The posterior mean spatial effect and its exceedance probability are shown in Figure 49. The estimated spatial variance proportion was 0.773, with interval $(0.378, 0.976)$, indicating that most residual heterogeneity was spatially structured after the measured county covariates were included.

The point process and areal formulations are connected. If a log-Gaussian Cox process is observed only through regional totals, integrating its intensity over each region leads to an areal count likelihood. Aggregation, however, discards locations within regions and can conceal fine scale clustering. Conversely, areal models may be preferable when the scientific unit is genuinely regional or when confidentiality prevents the release of exact locations.

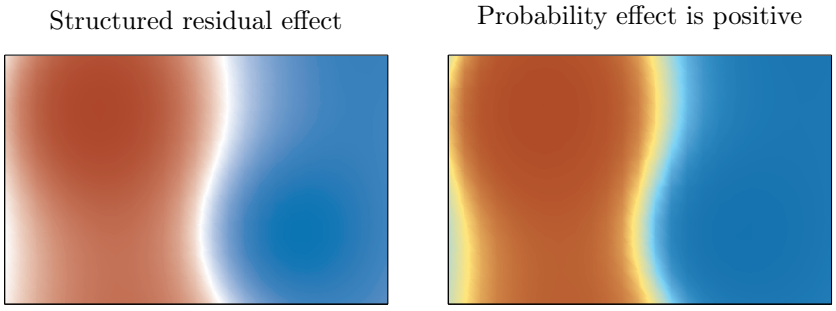

Figure 49

The original convolution model is due to Besag, York, and Mollié. Its modern scaled parametrization uses the penalized complexity prior principle described in Lecture 4.

Lecture 11 — Wednesday, March 27, 2019

Case-Control Sampling and Odds Ratios

A **cohort study** samples individuals according to exposure status and follows them to observe disease. It can estimate disease incidence in exposed and unexposed groups directly, but a rare disease may require a very large cohort and long follow up. A **case-control study** instead samples individuals according to disease status: cases have the disease and controls do not. Exposure is then compared retrospectively.

Let $Y = 1$ denote disease and $X = 1$ denote exposure. In the **source population**, write

$$p_1 = \mathbb{P}(X = 1 \mid Y = 1) \qquad \text{and} \qquad p_0 = \mathbb{P}(X = 1 \mid Y = 0).$$

The exposure odds ratio in a case-control table is

$$\mathrm{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}. \tag{11.1}$$

Bayes' rule gives the same quantity when disease odds are compared across exposure groups:

$$\begin{aligned} \frac{\mathbb{P}(Y=1 \mid X=1)/\mathbb{P}(Y=0 \mid X=1)}{\mathbb{P}(Y=1 \mid X=0)/\mathbb{P}(Y=0 \mid X=0)} &= \frac{\mathbb{P}(X=1 \mid Y=1)/\mathbb{P}(X=1 \mid Y=0)}{\mathbb{P}(X=0 \mid Y=1)/\mathbb{P}(X=0 \mid Y=0)} \\ &= \mathrm{OR}\,. \end{aligned} \tag{11.2}$$

This symmetry is the central reason that **retrospective sampling** is useful. The sampled proportions of cases and controls do not reproduce disease prevalence, but the cross product odds ratio remains meaningful.

Example 11.1 Doll and Hill compared smoking histories for lung cancer cases and hospital controls. Their male data, grouped by approximate daily cigarette consumption, were

Daily cigarettes	Cases	Controls	Odds ratio
None	2	27	1.0
1–4	24	38	8.5
5–14	208	242	11.6
15–24	196	201	13.2
25–49	174	118	19.9
50 or more	45	23	26.4

Within a category, the case to control ratio estimates the source-population disease odds up to a common sampling factor. Dividing by the corresponding ratio for nonsmokers eliminates that factor. The resulting odds ratios increase strongly with cigarette consumption.

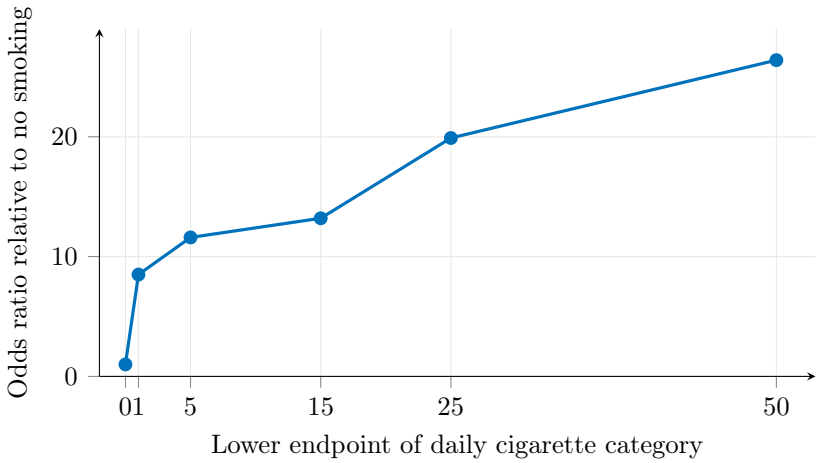

Figure 50

The dose-response pattern in the retrospective smoking odds ratios is visible in Figure 50.

The validity of this comparison depends on how individuals enter the study. Let $Z = 1$ indicate inclusion. Sampling may depend strongly on Y, because the numbers of cases and controls are

deliberately chosen. The crucial requirement is that, within each disease group, selection not depend on exposure or other predictors.

Theorem 11.2 Suppose that

$$\mathbb{P}(Z = 1 \mid Y, X) = \mathbb{P}(Z = 1 \mid Y).$$

Assume that both disease groups have positive sampling probabilities and that the disease odds are finite and positive at the covariate values being compared. Then the disease odds ratio is the same in the sampled and source populations.

Proof. Let $s_y = \mathbb{P}(Z = 1 \mid Y = y)$. Bayes' rule and the sampling assumption give

$$\begin{aligned} \frac{\mathbb{P}(Y = 1 \mid X = x, Z = 1)}{\mathbb{P}(Y = 0 \mid X = x, Z = 1)} &= \frac{\mathbb{P}(Z = 1 \mid Y = 1, X = x)\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Z = 1 \mid Y = 0, X = x)\mathbb{P}(Y = 0 \mid X = x)} \\ &= \frac{s_1}{s_0}\frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = 0 \mid X = x)}. \end{aligned} \tag{11.3}$$

When the odds at x_1 are divided by the odds at x_0, the sampling factor s_1/s_0 cancels. The two odds ratios are therefore equal. □

If the source population model is

$$\operatorname{logit} \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}, \tag{11.4}$$

then taking logarithms in (11.3) gives

$$\operatorname{logit} \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}, Z = 1) = \beta_0^{\star} + \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta} \qquad \text{and} \qquad \beta_0^{\star} = \beta_0 + \log\left(\frac{s_1}{s_0}\right). \tag{11.5}$$

Ordinary logistic regression fitted to an unmatched case-control sample therefore estimates the slope coefficients and their odds ratios, but its intercept is shifted. Without external prevalence or known sampling fractions, it cannot estimate absolute disease probabilities.

The assumption in Theorem 11.2 fails if exposure affects recruitment within case or control groups. If healthy smokers are less likely to participate, smokers are underrepresented among controls and the exposure odds ratio is biased upward. If smokers worried about disease are especially likely to volunteer as controls, the bias is downward. Recall error, inappropriate control selection, confounding, and differential diagnosis can create additional bias even when the algebraic sampling condition holds.

The odds ratio is not generally a risk ratio. If $q_x = \mathbb{P}(Y = 1 \mid X = x)$, then

$$\text{OR} = \frac{q_1/(1-q_1)}{q_0/(1-q_0)}.$$

When both risks are small, $1-q_1$ and $1-q_0$ are close to one, so the odds ratio approximates q_1/q_0. This **rare disease approximation** is unnecessary for the exact odds ratio identity in (11.2); it is needed only when an odds ratio is interpreted approximately as a risk ratio.

The retrospective odds ratio argument was formalized by Cornfield. The historical lung cancer study is available in the original report by Doll and Hill.

Matching, Conditional Likelihood, and Random Effects

An unmatched study of pediatric cervical spine injury compared 540 cases with 1060 randomly selected controls from seventeen hospitals. A logistic model included predisposing conditions, altered mental status, axial loading, several high-risk mechanisms, and clotheslining injury. A hospital random intercept allowed children treated at the same site to share unmeasured characteristics. The fitted model was

$$\operatorname{logit} \mathbb{P}(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}, U_j) = \beta_0 + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + U_j \qquad \text{and} \qquad U_j \sim \mathcal{N}(0, \sigma_U^2).$$

Its posterior summaries were

Predictor	Mean	2.5% quantile	97.5% quantile
Intercept	−1.08	−1.25	−0.91
Predisposing condition	1.84	0.74	3.09
Altered mental status	0.83	0.57	1.10
Axial loading	0.37	0.11	0.63
High-risk motor vehicle collision	0.53	0.23	0.83
High-risk collision with a vehicle	−0.49	−0.84	−0.15
High-risk diving	4.28	2.57	6.54
High-risk hanging	−5.35	−35.01	12.58
Clotheslining injury	1.42	0.53	2.35

Several predictors have clear conditional associations with injury. The extremely wide interval for high-risk hanging reflects sparse data or separation, so its large point estimate should not be interpreted substantively. The posterior mean hospital standard deviation was only 0.045, with interval $(0.002, 0.135)$, suggesting little residual hospital heterogeneity after adjustment.

Matching deliberately selects controls resembling each case on variables such as age, hospital, or treatment pathway. Let stratum i contain one case and $m_i - 1$ matched controls, and suppose that

$$\operatorname{logit} \mathbb{P}(Y_{ij} = 1 \mid \boldsymbol{x}_{ij}) = \alpha_i + \boldsymbol{x}_{ij}^{\mathsf{T}} \boldsymbol{\beta}. \tag{11.6}$$

Assume that the responses are conditionally independent within each stratum under this model. The nuisance intercept α_i absorbs all variables constant within the matched set, including the matching factors. Retrospective sampling changes this intercept but leaves $\boldsymbol{\beta}$ unchanged under the analogue of Theorem 11.2 for each stratum.

Condition on the event that exactly one member of stratum i is a case. If member k is that case, conditional independence in (11.6) gives

$$\begin{aligned} \mathbb{P}\left(Y_{ik} = 1 \;\middle|\; \sum_{j=1}^{m_i} Y_{ij} = 1, \mathbf{X}_i \right) &= \frac{\mathbb{P}(Y_{ik}=1) \prod_{\ell \neq k} \mathbb{P}(Y_{i\ell}=0)}{\sum_{h=1}^{m_i} \mathbb{P}(Y_{ih}=1) \prod_{\ell \neq h} \mathbb{P}(Y_{i\ell}=0)} \\ &= \frac{\exp\left(\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\beta}\right)}{\sum_{j=1}^{m_i} \exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta}\right)}. \end{aligned} \tag{11.7}$$

The stratum intercept cancels. If the observed case is indexed by $j = 1$, the **conditional logistic likelihood** is therefore

$$L_c(\boldsymbol{\beta}) = \prod_{i} \frac{\exp\left(\boldsymbol{x}_{i1}^{\mathsf{T}}\boldsymbol{\beta}\right)}{\sum_{j=1}^{m_i} \exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\beta}\right)} = \prod_{i} \left[\sum_{j=1}^{m_i} \exp\left((\boldsymbol{x}_{ij} - \boldsymbol{x}_{i1})^{\mathsf{T}} \boldsymbol{\beta} \right) \right]^{-1}. \tag{11.8}$$

Only covariate contrasts within a stratum contribute to (11.8). Consequently, the effect of a matching variable cannot be estimated, and a covariate with little variation within matched sets will have a large standard error even if it varies substantially across the full sample.

In the cervical spine study, each case receiving emergency medical services was matched to as many as two controls of similar age who also received such care. Conditional logistic regression produced the following estimates.

Predictor	Coefficient	Odds ratio	Standard error
Predisposing condition	0.00	1.00	0.99
Altered mental status	1.09	2.96	0.19
Axial loading	0.57	1.77	0.20
High-risk motor vehicle collision	0.81	2.24	0.20
High-risk collision with a vehicle	−0.42	0.66	0.22
High-risk diving	3.14	23.08	1.05
Clotheslining injury	1.78	5.94	0.91

The controls were children with trauma rather than a random sample of healthy children. Thus, these odds ratios distinguish cervical spine injury cases from comparable injured children who received emergency care; they do not compare cases with the general pediatric population. The estimand follows from the control sampling mechanism.

Additional random effects can be difficult in a matched design. Conditioning removes one intercept for every matched set, but hospital, clinician, or geographic effects may cross those sets. The likelihood then combines conditional logistic terms with another latent distribution. Markov chain Monte Carlo, adaptive quadrature, or specialized approximations may be needed, and identifiability depends on sufficient replication across both strata and clusters.

Data from cases alone pose a sharper interpretive problem. A database of deaths involving police contains fatal events but not the full set of police encounters. It therefore cannot directly estimate

$$\mathbb{P}(\text{death} \mid \text{encounter}, \text{armed status}, \text{covariates}).$$

Nevertheless, a transparent generative model shows which relative quantities can be recovered under additional assumptions.

Index race or ethnicity by r, state by s, and other covariates by $\boldsymbol{x}$. Let $P_{rs\boldsymbol{x}}$ be the population size, $\lambda_{rs\boldsymbol{x}}$ the police encounter rate, and $\rho_{rs\boldsymbol{x}}$ the probability that a person is armed during an encounter. If $\gamma_{ars\boldsymbol{x}}$ is the probability that an encounter is fatal when armed status is a, then the armed and unarmed fatal counts have means

$$\mathbb{E}[Y_{1rs\boldsymbol{x}}] = P_{rs\boldsymbol{x}}\lambda_{rs\boldsymbol{x}}\rho_{rs\boldsymbol{x}}\gamma_{1rs\boldsymbol{x}}, \tag{11.9}$$

$$\mathbb{E}[Y_{0rs\boldsymbol{x}}] = P_{rs\boldsymbol{x}}\lambda_{rs\boldsymbol{x}}(1-\rho_{rs\boldsymbol{x}})\gamma_{0rs\boldsymbol{x}}. \tag{11.10}$$

Conditional on a recorded fatality, the common population and encounter rate terms cancel:

$$\frac{\mathbb{P}(A=1 \mid \text{fatality}, r, s, \boldsymbol{x})}{\mathbb{P}(A=0 \mid \text{fatality}, r, s, \boldsymbol{x})} = \frac{\gamma_{1rs\boldsymbol{x}}}{\gamma_{0rs\boldsymbol{x}}}\frac{\rho_{rs\boldsymbol{x}}}{1-\rho_{rs\boldsymbol{x}}}. \tag{11.11}$$

The observed armed status odds therefore combine two mechanisms: how often people are armed during encounters and how fatality probabilities differ between armed and unarmed encounters. Neither mechanism is separately identified by fatalities alone.

To obtain a logistic regression, suppose that armed status odds during an encounter do not depend on race or ethnicity after state and other covariates are held fixed:

$$\frac{\rho_{rs\boldsymbol{x}}}{1-\rho_{rs\boldsymbol{x}}} = W_s \delta_{\boldsymbol{x}}.$$

Also suppose that the fatality probabilities factor as

$$\gamma_{ars\boldsymbol{x}} = \nu_{ar}\alpha_{a\boldsymbol{x}}V_{sa}.$$

Equation (11.11) then becomes

$$\operatorname{logit}\mathbb{P}(A_i = 1 \mid \text{fatality}, r_i, s_i, \boldsymbol{x}_i) = \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\beta} + U_{s_i},$$

where

$$U_s = \log\left(\frac{W_s V_{s1}}{V_{s0}}\right).$$

The state effect conflates state differences in armed encounters with state differences in the relative fatality probabilities for armed and unarmed encounters.

An analysis of recorded fatalities among White, Black, and Hispanic or Latino individuals estimated the following coefficients, with White males as the reference group.

Quantity	Median	2.5% quantile	97.5% quantile
Intercept	1.28	1.08	1.49
Black	−0.24	−0.52	0.03
Hispanic or Latino	−0.18	−0.53	0.18
Female	−1.10	−1.57	−0.62
State standard deviation	0.31	0.15	0.51

The response is armed status among recorded fatalities. A negative race coefficient is not, by itself, an estimate of higher fatality risk while unarmed. That interpretation would additionally require comparable armed encounter fatality probabilities across groups and the assumed race invariance of armed status during encounters. The female coefficient likewise combines sex differences in armed status with sex differences in fatality probabilities.

The state effects display genuine heterogeneity among states in the conditional armed status odds, but their decomposition is not identifiable from the fatality data. External information about encounter counts, gun carrying, or reporting coverage would be needed to distinguish the possible mechanisms.

Finally, the derivation assumes that recorded fatalities are representative within the relevant armed status, race, state, and covariate groups. If the probability that a fatality is recorded varies by armed status and race, the odds in (11.11) acquire another unidentified selection factor. Sensitivity analysis should vary such factors rather than treat one identifying assumption as fact.

Retrospective designs preserve odds ratios under sampling based on outcomes only when their sampling assumptions hold. Absolute risks require population information, matching defines the comparison set, and analyses of cases alone identify only combinations of encounter and outcome processes.

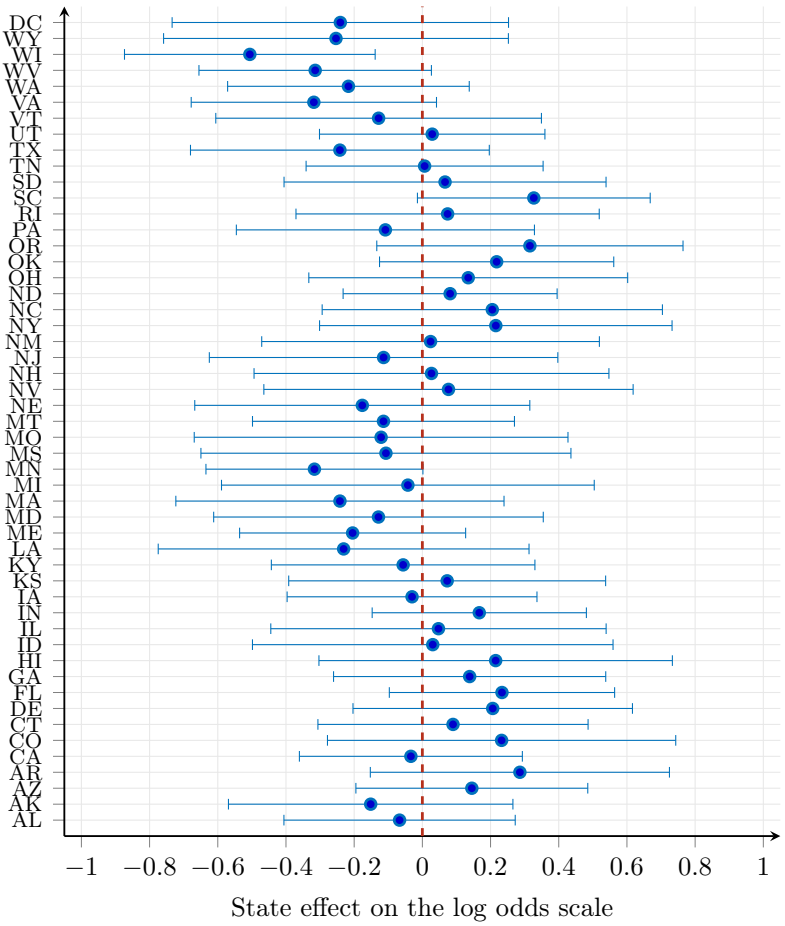

Figure 51

The fitted heterogeneity among states in the conditional odds of being armed is displayed in Figure 51.

Appendix: Lecture and Tutorial Index