Stat/Biostat 571 Statistical Methodology: Regression Models for Dependent Data

Jon Wakefield

Departments of Statistics and Biostatistics, UW.

Lectures: Monday/Wednesday/Friday 1.30–2.20, T635.

Coursework: weekly (30%). Examination at mid-term (30%) and final (40%).

Office Hours:

Jon: Monday 2.30–3.20 and Wednesday 2.30–3.30, both in Health Sciences, F664. Or by appointment (Email: `jonno@u.washington.edu`, Phone: 616-6292).

Kent Koprowicz (`kentk@u`)

STAT/BIOSTAT 578 Data Analysis, strongly recommended for Applied Exam. This course teaches methods, not data analysis.

Computing will be carried out using `R/Splus` and `WinBUGS`.

Class website: `http://courses.washington.edu/b571/`

Textbooks:

Main Texts

Diggle, P.J., Heagerty, P., Liang, K.-Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data, Second Edition*. Oxford University Press: this text is closest to the material covered in the course.

Fitzmaurice, G.M., Laird, N.M. and Ware, J.H. (2004). *Applied Longitudinal Analysis*, Wiley.

Gelfand, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, CRC Press.

Hand, D. and Crowder, M.J. (1996). *Practical Longitudinal Data Analysis*, CRC Press.

Pinheiro, J. and Bates, D.G. (2000). *Mixed-Effects Models in S and S-PLUS*, Springer-Verlag,

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer-Verlag.

Background Texts

Davison, A.C. (2003). *Statistical Models*. Cambridge University Press.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*, Wiley.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models, Second Edition*, CRC Press.

COURSE OUTLINE

Chapter 1 Revisited

Motivating Datasets; Benefits and Challenges of Dependent Data; Marginal versus Conditional Modeling.

Chapter 8: Linear Models

Linear Mixed Effects Models; Frequentist and Bayesian Inference; Equivalence of Marginal and Conditional Modeling.

Chapter 9: General Regression Models

Generalized Linear Mixed Models; Frequentist and Bayesian Inference; Non-equivalence of Marginal and Conditional Modeling.

Chapter 10: Binary Data Models

Modeling the covariance structure. Mixed Effects approach.

Chapter 11: Model Selection/Formulation

Types of analysis: descriptive, confirmatory, predictive. Causality and confounding.

CHAPTER 1: OVERVIEW

Recall: in a *regression analysis* we model a response, Y, as a function of covariates, $\boldsymbol{x}$.

In 570 we considered situations in which responses are *conditionally independent*, that is

$$\begin{aligned} p(Y_1,...,Y_n|\boldsymbol{\beta},\boldsymbol{x}) &= p(Y_1|\boldsymbol{\beta},\boldsymbol{x}_1)\times p(Y_2|Y_1,\boldsymbol{\beta},\boldsymbol{x}_2)\times ...\\ &\times\ p(Y_n|Y_1,...,Y_{n-1},\boldsymbol{\beta},\boldsymbol{x}_n)\\ &= p(Y_1|\boldsymbol{\beta},\boldsymbol{x}_1)\times p(Y_2|\boldsymbol{\beta},\boldsymbol{x}_2)\times ...\\ &\times\ p(Y_n|\boldsymbol{\beta},\boldsymbol{x}_n)\end{aligned}$$

so that observations are independent *given* parameters $\boldsymbol{\beta}$ and covariates $\boldsymbol{x}_1,...,\boldsymbol{x}_n$.

In general, $Y_1,...,Y_n$ are *never* independent. For example, suppose

$$\text{E}[Y_i|\mu,\sigma^2]=\mu,\quad \text{var}(Y_i|\mu,\sigma^2)=\sigma^2,$$

$i=1,2$ and $\text{cov}(Y_1,Y_2|\mu,\sigma^2)=0$. Then if we are told y_1, this will change the way we think about y_2 so that

$$p(Y_2|Y_1)\neq p(Y_2),$$

and the observations are not independent, however

$$p(Y_2|Y_1,\mu,\sigma^2)=p(Y_2|\mu,\sigma^2),$$

so that we have conditional independence.

Motivating Examples

We distinguish between dependence induced by missing covariates, and that due to contagion (for example, in an infectious disease context) – we will not consider the latter.

One theme of the course will be modeling dependence in the *residuals*, that is, after we have controlled for covariates.

The obvious situations in which we would expect dependence is in data collected over time or space.

Example 1: Dental growth data

Table 1 records dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure in 11 girls and 16 boys at the ages of 8, 10, 12 and 14 years.

Here we have an example of *repeated measures* or *longitudinal* data.

Figure 1 plots these data and we see that dental growth for each child increases in an approximately linear fashion.

One common aim of such studies is to identify the *within-individual* and *between-individual* sources of variability.

Individual	Years			
	8	10	12	14
Girls				
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
Boys				
1	26	25	29	31
2	21.5	22.5	23	26.5
3	23	22.5	24	27.5
4	25.5	27.5	26.5	27
5	20	23.5	22.5	26
6	24.5	25.5	27	28.5
7	22	22	24.5	26.5
8	24	21.5	24.5	25.5
9	23	20.5	31	26
10	27.5	28	31	31.5
11	23	23	23.5	25
12	21.5	23.5	24	28
13	17	24.5	26	29.5
14	22.5	25.5	25.5	26
15	23	24.5	26	30
16	22	21.5	23.5	25

Table 1: Dental growth data for girls and boys.

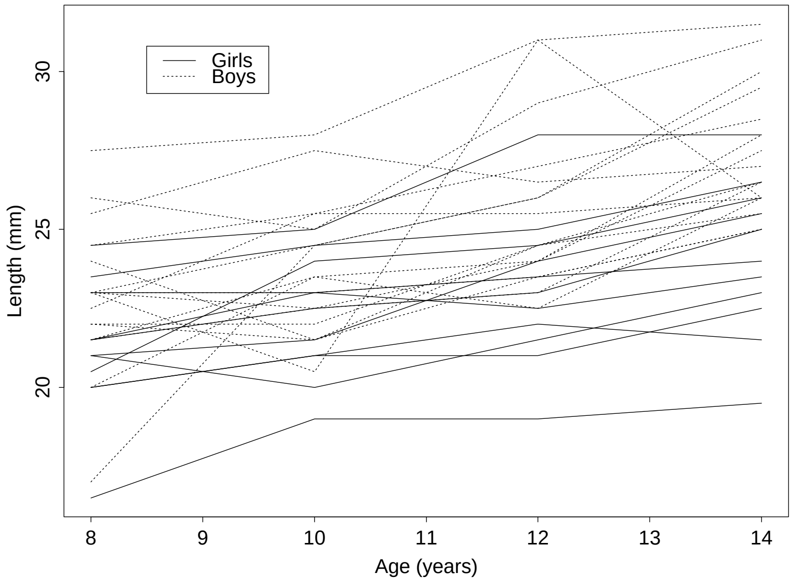

Figure 1: Dental growth data for girls and boys.

Inference

We may be interested in characterizing:

1. the *average* growth curve, or
2. the growth for a *particular* child.

Two types of analysis that will be distinguished are *marginal* and *conditional*. The former is designed for questions of type 1, and the latter for type 2.

Even if the question of interest is of type 1, we still have to acknowledge the dependence of responses on the same individual – we do not have 11×4 independent observations on girls and 16×4 independent observations on boys but rather 11 and 16 *sets* of observations on girls and boys.

For either question of interest ignoring the dependence leads to incorrect standard errors and confidence interval coverage.

A marginal approach to modeling specifies the moments of the data only, while in a conditional approach the responses of specific individuals are modeled.

Models

First question is: why not just analyze the data from each child separately? Possible but we wouldn't be able to make formal statements about:

- The average growth rate of teeth for a girl in the age range 8–14 years.
- The between-girl variability in growth rates.

The totality of data on girls may also aid in the estimation of the growth rate for a particular girl – becomes more critical as the number of observations per child decreases. For example, in an extreme case, suppose a particular girl has only one measurement?

At the other extreme we could fit a single curve to the data from all of the girl's data together. The problem with this is that we do not have independent observations, and what if we are interested in inference for a particular child?

Example 2: Spatial Studies

Dependent data may result from studies with a significant spatial component.

Split Plot Data

Example: Three varieties of oats, four nitrogen concentrations.

Agricultural land was grouped into six blocks, each with three plots, and with each plot further sub-divided into four sub-plots. Within each subplot a combination of oats and nitrogen was planted. Hence we have $6 \times 3 \times 4 = 72$ observations.

We would expect observations within the same sub-plot to be correlated.

Disease Mapping Data

We have a set of counts of disease, and population sizes for a set of m areas that partition a study area. We expect rates of disease in two areas to display greater correlation if those areas are geographically close.

Aims:

- Simple description – a visual summary of geographical risk.
- Provide estimates of risk by area to inform public health resource allocation.
- Give clues to etiology via informal examination of maps with exposure maps, components of spatial versus non-spatial residual variability may also provide clues to source of variability (e.g. environmental exposures usually have spatial structure). The formal examination is carried out via spatial regression.
- Provide a context within which specific studies may be placed.

Example: Lung and Brain cancer in the North-West of England

This study will be used as an illustration of smoothing techniques using a variety of hierarchical models.

Two tumors were chosen to contrast mapping techniques for relatively non-rare (lung), and relatively rare (brain) cancers.

The absence of information on smoking means that for lung cancer in particular the analysis should be viewed as illustrative only (since a large fraction of the residual variability would disappear if smoking information were included).

Study details:

- Study period is 1981–1991.
- Incidence data by postcode, but the analysis is carried out at the ward level of which there are 144 in the study region. For brain cancer the median number of cases per ward over the 11 year period is 6 with a range of 0 to 17. For lung the median number is 20 with range 0–60.
- Expected counts were based on ward-level populations from the 1991 census, by 5-year age bands and sex.

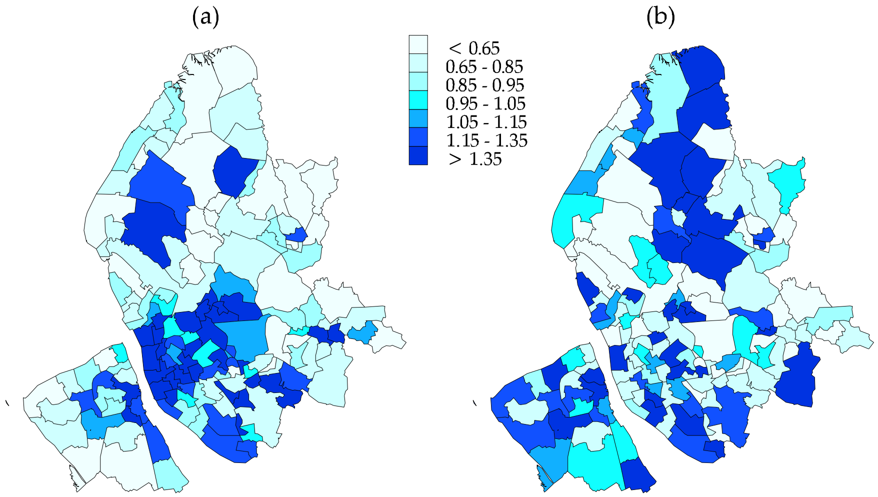

Figure 2: SIRs for (a) lung cancer, and (b) brain cancer.

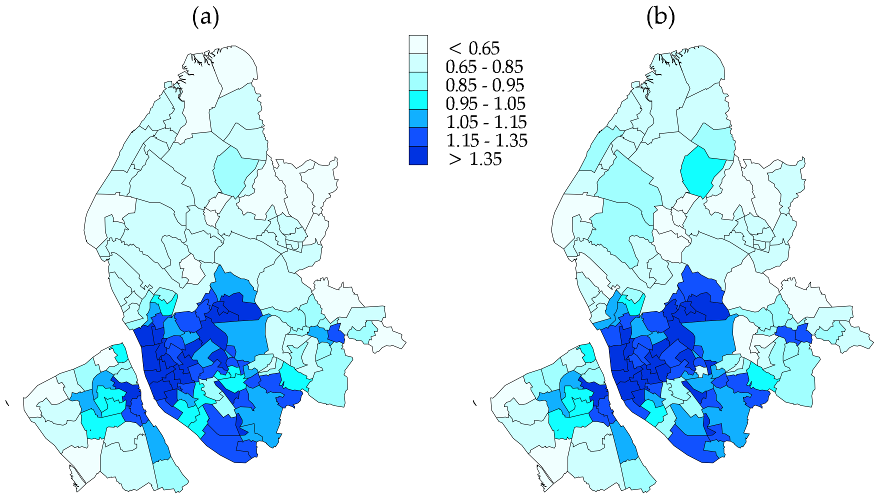

Figure 3: Smoothed SIRs for lung cancer under (a) a conditional spatial model, and (b) a marginal spatial model.

Notice that the smoothed area-level relative risk estimates are not dramatically different from the raw versions in Figure 2(a) – the large number of cases here mean that the raw SIRs are relatively stable.

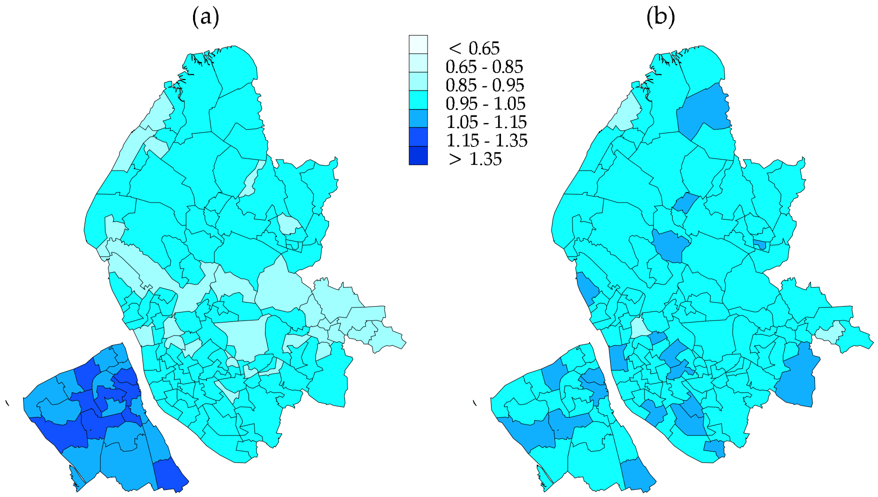

Figure 4: Smoothed SIRs for brain cancer under (a) a conditional spatial model, an d (b) a marginal spatial model.

In this case we see a much greater smoothing of the estimates as compared to the raw relative risks in Figure 2(b).

Design Implications of a Longitudinal Study

To examine a the implications of carrying out a longitudinal study, as compared to a cross-sectional study, we consider a very simple situation in which we wish to compare two treatments, coded as -1 and +1, and we have a linear model.

Cross-Sectional Study:

A single measurement is taken on each of $m = 4$ individuals where

$$Y_{i1} = \beta_0 + \beta_1 x_{i1} + \epsilon_{i1},$$

$i = 1, ..., m = 4$, ϵ_{i1} iid with $\text{var}(\epsilon_{i1}) = \sigma^2$ and $x_{11} = -1, x_{21} = -1, x_{31} = 1, x_{41} = 1.$

Note: $\text{E}[Y_1|x = 1] - \text{E}[Y_1|x = -1] = 2\beta_1.$

In lectures will show that

$$\widehat{\beta}_0^c = \frac{\sum_{i=1}^4 Y_{i1}}{4}, \quad \widehat{\beta}_1^c = \frac{Y_{31} + Y_{41} - (Y_{11} + Y_{21})}{4},$$

and

$$\text{var}(\widehat{\beta}_0^c) = \text{var}(\widehat{\beta}_1^c) = \frac{\sigma^2}{4}.$$

Longitudinal Study:

We assume the model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + \delta_{ij},$$

with α_i and δ_{ij} independent and with $\text{var}(\alpha_i) = \sigma_\alpha^2$, $\text{var}(\delta_{ij}) = \sigma_\delta^2$. We therefore have marginally:

$$\text{var}(Y_{ij}|\beta_0, \beta_1) = \sigma_\alpha^2 + \sigma_\delta^2 = \sigma^2,$$

and

$$\text{cov}(Y_{i1}, Y_{i2}) = \sigma_\alpha^2.$$

We let $\rho = \sigma_\alpha^2/\sigma^2$, represent the correlation on observations on the same individual.

We consider two situations, both with two observations on two individuals:

Constant treatment for each individual:

$$x_{11} = x_{12} = -1, \quad x_{21} = x_{22} = 1.$$

Changing treatment for each individual:

$$x_{11} = x_{22} = 1, \quad x_{12} = x_{21} = -1.$$

Using Generalized Least Squares we have

$$\widehat{\boldsymbol{\beta}}^l = (\boldsymbol{x}^{\text{T}}\boldsymbol{R}^{-1}\boldsymbol{x})^{-1}\boldsymbol{x}^{\text{T}}\boldsymbol{R}^{-1}\boldsymbol{Y},$$

and

$$\text{var}(\widehat{\boldsymbol{\beta}}^l) = (\boldsymbol{x}^{\text{T}}\boldsymbol{R}^{-1}\boldsymbol{x})^{-1}\sigma^2,$$

where

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}.$$

In lectures we will show that

$$\text{var}(\widehat{\beta}_1^l) = \frac{\sigma^2(1-\rho^2)}{4 - 2\rho(x_{11}x_{12} + x_{21}x_{22})}.$$

The *efficiency* e is given by

$$e = \frac{\text{var}(\widehat{\beta}_1^l)}{\text{var}(\widehat{\beta}_1^c)} = \frac{(1-\rho^2)}{1-\rho(x_{11}x_{12}+x_{21}x_{22})/2}.$$

Usually we have $\rho > 0$.

For the constant treatment longitudinal study

$$e = 1 + \rho,$$

so that the cross-sectional study is preferable since we have lost information due to the correlation.

For the changing treatment longitudinal study

$$e = 1 - \rho,$$

so that the longitudinal study is more efficient, because each individual is acting as their own control, that is, we are making within-individual comparisons.

If $\rho = 0$ the designs have the same efficiency.

Introduction

We consider the situation in which we have a vector of responses, $\boldsymbol{Y}_i = (Y_{i1}, ..., Y_{in_i})$, for the $i-$th unit, $i = 1, ..., m$, with the mean for $\boldsymbol{Y}_i$ being linear in a $(k+1) \times 1$ vector of covariates $\boldsymbol{x}_i$.

We assume that the responses on different units are independent, but that there is dependence between observations on the same unit.

In a *balanced* data set all units have the same number of observations, and are observed at a common set of occasions, so that $n_i = n$. In an *unbalanced* data set this is not the case.