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, T478.

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

Office Hours:

Jon: Monday 2.30–3.20 (Statistics, Padelford, 616-9388) and Wednesday 2.30–3.30 (Biostatistics, Health Sciences, 616-6292). Or by appointment (`jonno@u.washington.edu`).

TA's: Liz Thomas (`lizt@u`), Ben French (`bcf@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` 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.

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

Gelman, 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: Revision

Motivating Datasets; Benefits and Challenges of Dependent Data; Marginal versus Conditional Modeling. Sandwich Estimation; Ordinary and Weighted Least Squares.

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

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

$i = 1, 2$ and $\mathrm{cov}(Y_1, Y_2|\mu, \sigma^2) = 0$. 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 *residual* dependence, i.e. after we have controlled for covariates.

The obvious situations in which we would expect dependence is in data collected over time or space (but lots of others possible, e.g. families).

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.

Girls	8	10	12	14
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	8	10	12	14
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

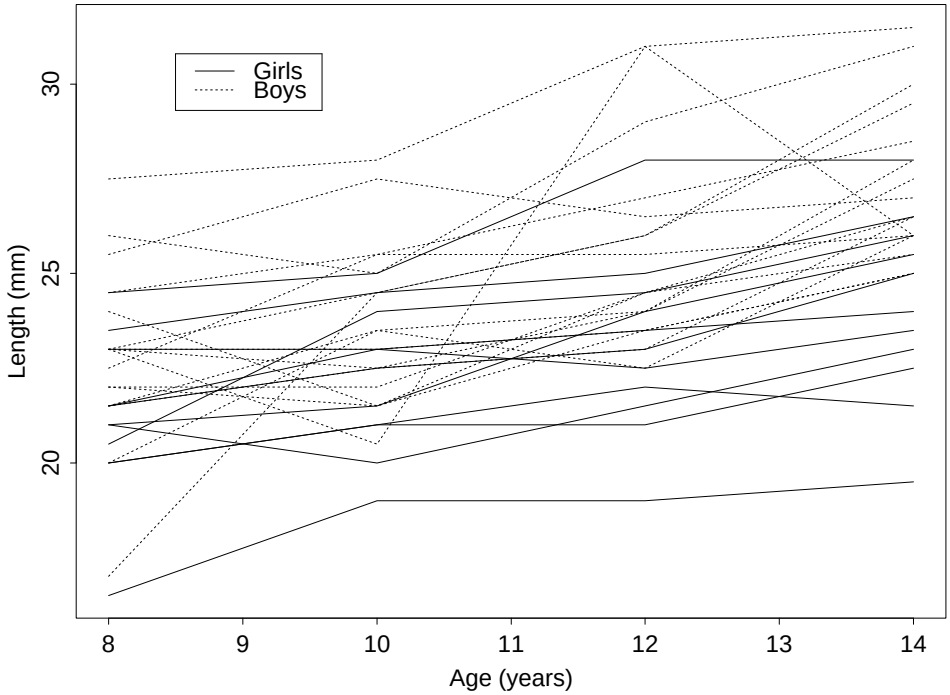

Figure 1: Dental growth data for girls and boys.

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 may be used for both types, but requires more assumptions.

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 Data

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 block to be correlated.

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

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.
- Standardized Incidence Rates (SIRs) for area i are calculated as Y_i/E_i where Y_i and E_i are observed and expected cases.

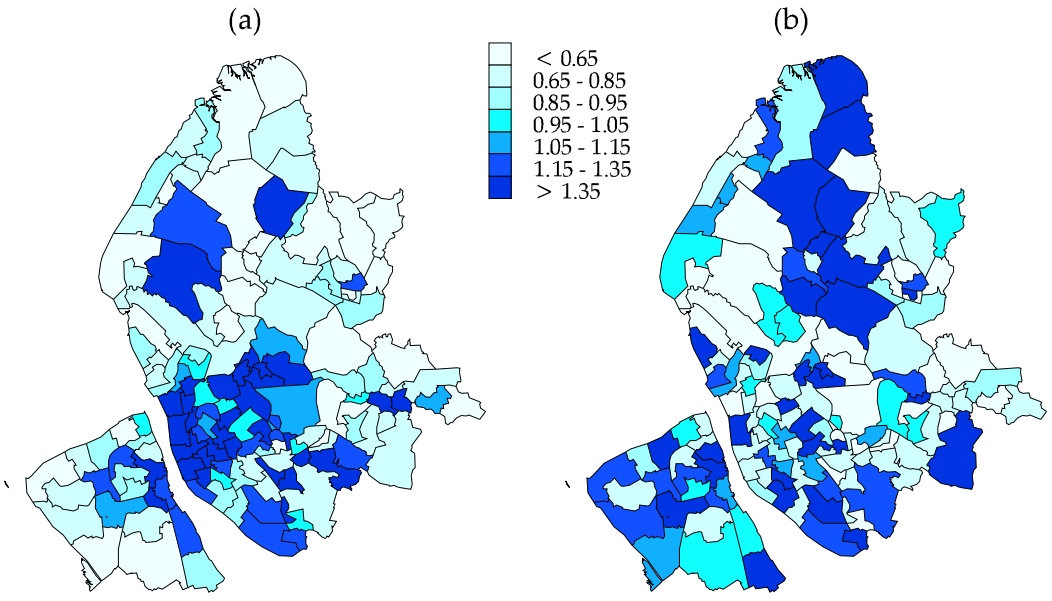

Figure 2: SIRs for (a) lung cancer, and (b) brain cancer.

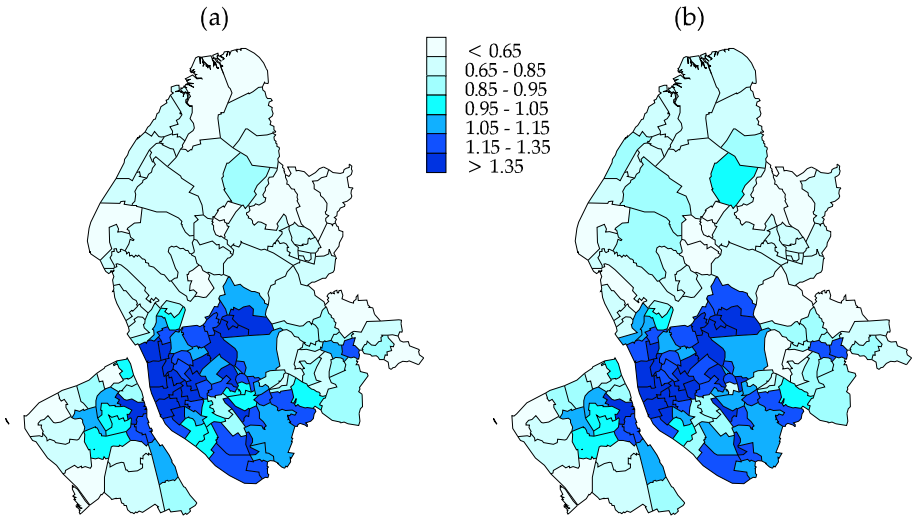

Figure 3: Smoothed SIRs for lung cancer under two spatial models.

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.

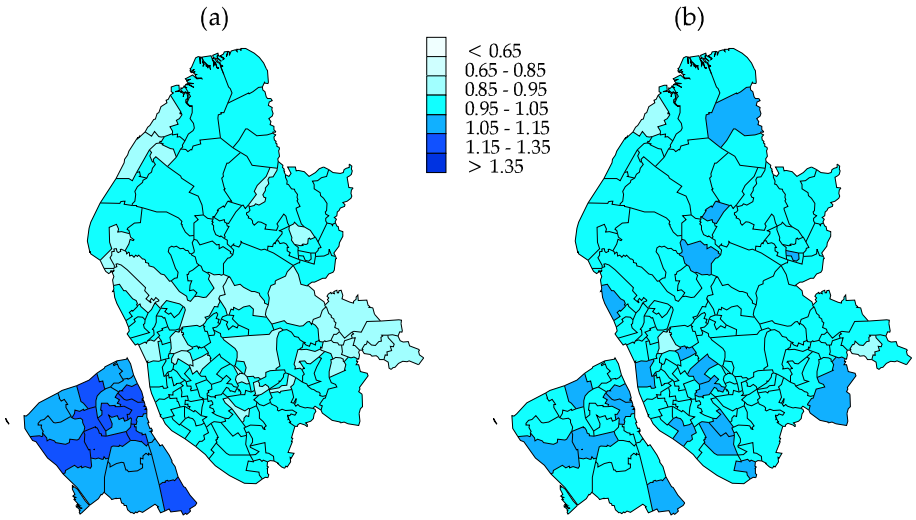

Figure 4: Smoothed SIRs for brain cancer under two spatial models.

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

Revision Material

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$, represent n observations from a distribution indexed by a p-dimensional parameter $\boldsymbol{\theta}$, with $\text{cov}(Y_i, Y_j \mid \boldsymbol{\theta}) = 0$, $i \neq j$.

In the following, for ease of presentation, we assume that Y_i, $i = 1, ..., n$ are independent and identically distributed (i.i.d.).

An *estimating function* is a function

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) \tag{1}$$

of the same dimension as $\boldsymbol{\theta}$ for which

$$\text{E}[\boldsymbol{G}_n(\boldsymbol{\theta})] = \boldsymbol{0} \tag{2}$$

for all $\boldsymbol{\theta}$. The estimating function $\boldsymbol{G}_n(\boldsymbol{\theta})$ is a random variable because it is a function of $\boldsymbol{Y}$.

The corresponding *estimating equation* that defines the estimator $\widehat{\boldsymbol{\theta}}_n$ has the form

$$\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\widehat{\boldsymbol{\theta}}_n, Y_i) = \boldsymbol{0}. \tag{3}$$

Result: Suppose that $\widehat{\boldsymbol{\theta}}_n$ is a solution to the estimating equation

$$\boldsymbol{G}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{G}(\boldsymbol{\theta}, Y_i) = \boldsymbol{0},$$

i.e. $\boldsymbol{G}_n(\widehat{\boldsymbol{\theta}}_n) = \boldsymbol{0}$. Then $\widehat{\boldsymbol{\theta}}_n \rightarrow_p \boldsymbol{\theta}$ (consistency) and

$$\sqrt{n}\ (\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \rightarrow_d \text{N}_p(\boldsymbol{0}, \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{\text{T}-1}) \tag{4}$$

(asymptotic normality) where

$$\boldsymbol{A} \quad = \quad \boldsymbol{A}(\boldsymbol{\theta}) = \text{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{G}(\boldsymbol{\theta}, Y)\right]$$

and

$$\boldsymbol{B} \quad = \quad \boldsymbol{B}(\boldsymbol{\theta}) = \text{E}[\boldsymbol{G}(\boldsymbol{\theta}, Y)\boldsymbol{G}(\boldsymbol{\theta}, Y)^{\text{T}}] = \text{cov}\{\boldsymbol{G}(\boldsymbol{\theta}, Y)\}.$$

The form of the variance in (4) has lead to it being named a **sandwich estimator**.

Least Squares Estimation

For the ordinary least squares/maximum likelihood estimator $\widehat{\boldsymbol{\beta}} = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}\boldsymbol{Y}$ with

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2$$

if $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\boldsymbol{I}$.

Suppose that $\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x}) = \sigma^2\boldsymbol{V}$ so that the model from which the estimator was derived was incorrect.

Then the estimator is still unbiased but the appropriate variance estimator is

$$\begin{aligned} \mathrm{var}(\widehat{\boldsymbol{\beta}}) &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y} \mid \boldsymbol{x})\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1} \\ &= (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2 \qquad (5) \end{aligned}$$

Expression (5) can also be derived directly from the estimating function

$$\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}),$$

from which we know that

$$(\boldsymbol{A}_n^{-1}\boldsymbol{B}_n\boldsymbol{A}_n^{\mathrm{T}\,-1})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;\rightarrow_d\; \mathrm{N}_{k+1}(\boldsymbol{0}, \boldsymbol{I}),$$

where

$$\boldsymbol{B}_n = \mathrm{var}(\boldsymbol{G}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}\sigma^2$$

and

$$\boldsymbol{A}_n = \mathrm{E}\left[\frac{\partial \boldsymbol{G}}{\partial \boldsymbol{\beta}}\right] = -\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x},$$

to give

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{x}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\sigma^2.$$

We still need to know $\boldsymbol{V}$ though.

Sandwich estimator with uncorrelated errors

We relax the constant variance assumptions. Consider the estimating function

$$\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}).$$

The "bread" of the sandwich, $\boldsymbol{A}^{-1}$, remains unchanged since $\boldsymbol{A}$ does not depend on Y.

The "filling" becomes

$$\boldsymbol{B} = \mathrm{var}\{\boldsymbol{G}\} = \boldsymbol{x}^{\mathrm{T}}\mathrm{var}(\boldsymbol{Y})\boldsymbol{x} = \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i, \tag{6}$$

where $\sigma_i^2 = \mathrm{var}(Y_i)$ and we have assumed that the data are uncorrelated.

Unfortunately σ_i^2 is unknown – we now discuss various estimation methods.

An obvious estimator is given by

$$\widehat{\boldsymbol{B}}_n = \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2, \tag{7}$$

and its use provides a consistent estimator of (6), if the data are uncorrelated.

For linear regression the estimator

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{\sigma}_i^2,$$

is downwardly biased, with bias $-p\sigma^2/n$.

The sandwich estimator is therefore also downwardly biased.

Using

$$\widetilde{\sigma}_i^2 = \frac{n}{n-k-1}(Y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}})^2 \tag{8}$$

provides a simple correction, but in general the estimator of the variance has finite bias since the bias in $\widehat{\sigma}^2$ changes as a function of the design points $\boldsymbol{x}_i$ – various corrections have been suggestions (see Kauermann and Carroll, 2001, JASA).