# BIOST / STAT 571

Winter 2022

Part 1: Introduction to Correlated Data

# Dependent, Correlated, Clustered, Longitudinal, Multivariate Data

- Longitudinal data
  - Collecting data on subjects over time
- Hierarchical data
  - Observations exist within a hierarchy
- Spatial data
  - Observations have correlation due to spatial characteristics
- Multivariate data
  - Measuring multiple characteristics of a subject as outcome
  - Repeated measures of an assay
  - Gene expression profiling experiment
  - Microbiome experiment

# Introduction to Correlated Data

## 1. Clustered data

Observations within the same cluster are likely to be correlated .

- Longitudinal study

| Subject | time |
|---|---|
| 1 | x x x |
| 2 | x x x |
| … | |
| m | x x x x x |

e.g. Outcome Y= Infection (Y/N)

# of seizures over time

CD4 counts over time

- Familial study

| Family | Member | | |
|---|---|---|---|
| 1 | x | x | x |
| 2 | x | x | |
| … | | | |
| m | x | x | |

e.g.

Outcome Y = Breast cancer (Y/N)
Time to death

- Toxicology

| Litter | Offspring (Rat) |
|---|---|
| 1 | x x x |
| 2 | x x |
| … | |
| m | x x x |

e.g.

Outcome Y = Alive /Dead
Birth weight
Normal / Malformation /Dead

## 2. **Hierarchical Data**

Observations from the same hierarchy are likely to be correlated.

- Survey Sampling

| state | county | household |
|---|---|---|
| 1 | 1 | x x x |
| | 2 | x x x x |
| | 3 | x x |
| 2 | 1 | x x x x x |
| | 2 | x x x |
| …. | | |

e.g.

Outcome Y = income

- Clinical trial

| center | physician | patient |
|---|---|---|
| 1 | 1 | x x |
| | 2 | x x x x |
| | 3 | x x x |
| 2 | 1 | x x x x x |
| | 2 | x x x |
| | 3 | x x |
| | 4 | x x x |
| …. | | |

e.g.

Outcome Y = heart disease (Y/N)

- Group-randomized Intervention Study

| school | class | student |
|---|---|---|
| 1 | 1 | x x x |
| | 2 | x x x x |
| | 3 | x x |
| | 4 | x x x x |
| 2 | 1 | x x |
| | 2 | x x |
| | 3 | x x x |
| …. | | |

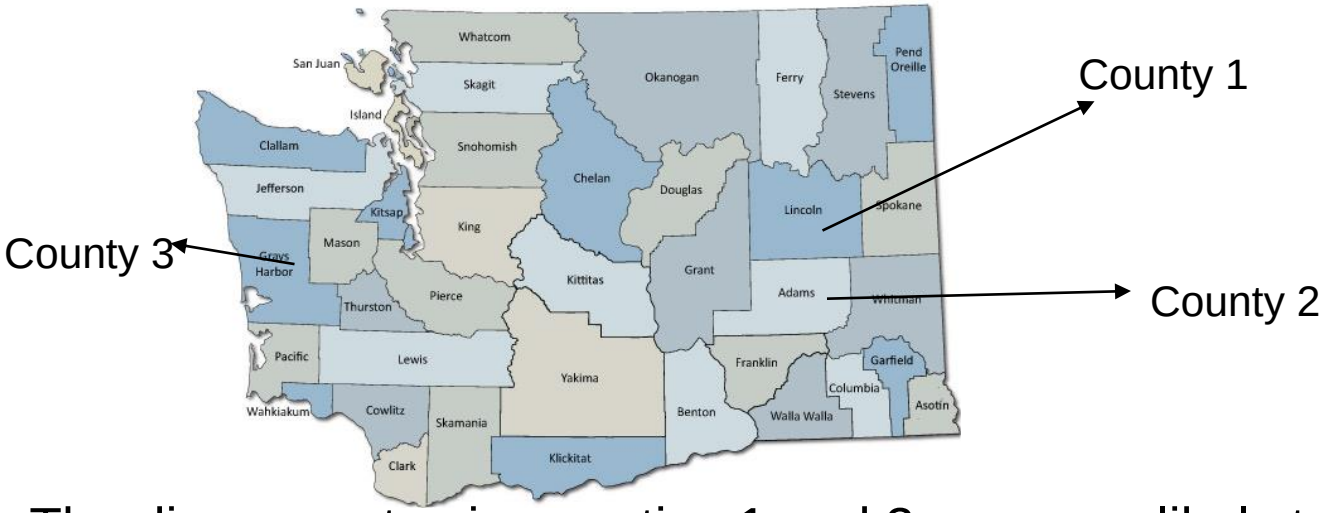
e.g.

Outcome Y = smoking

# 3. **Spatial Data**

Observations are correlated due to spatial “proximity.”

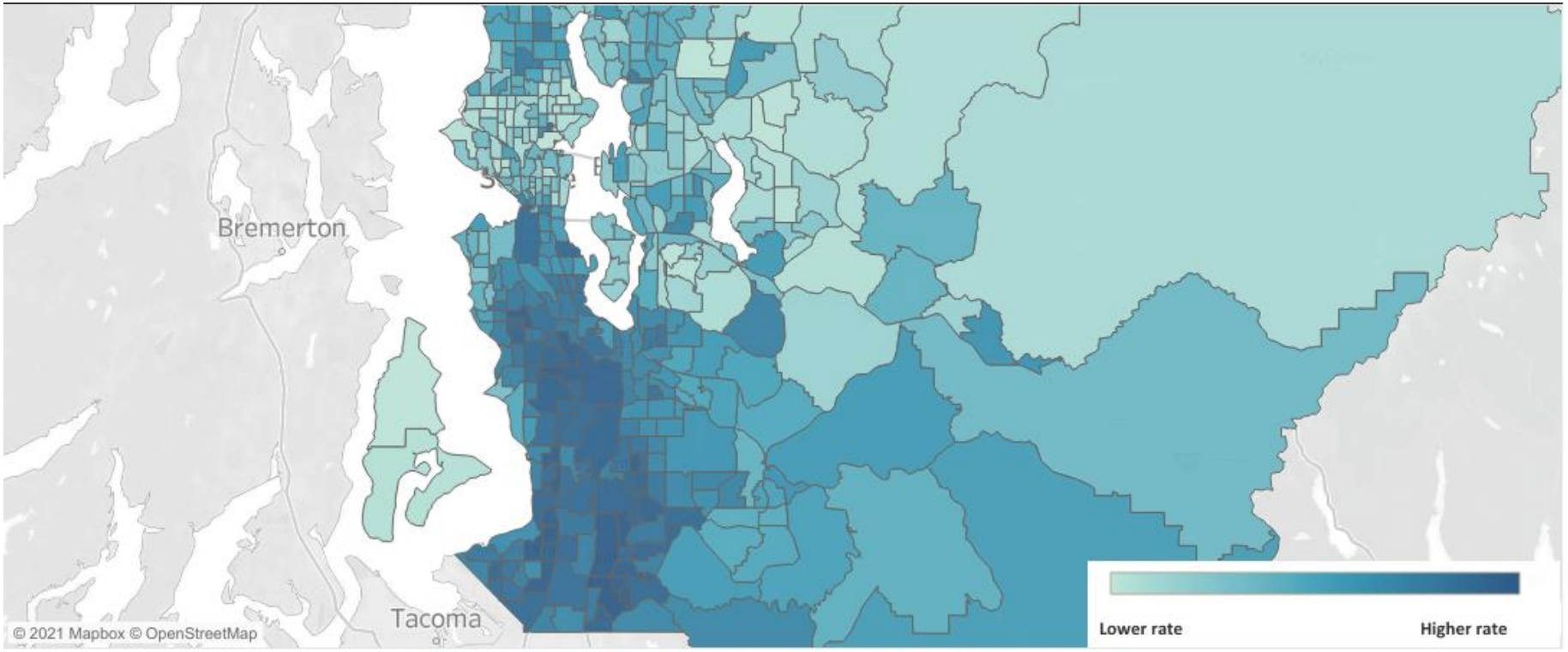
- Disease mapping




- The disease rates in counties 1 and 2 are more likely to be similar compared to the disease rates in counties 1 and 3.
- Y = # of disease cases (asthma cases)/10,000

  X = exposure (PM2.5)

- COVID case rates in King County




Overall King County rate of positive cases: 29.0 per 1,000 residents

- Nearby census districts have more similar rates

- Ecological Study

  To study the spatial distribution of a chemical and its relationship with the location of an industry site.

  - For example, the University of Michigan Dioxin Exposure Study.
  - Y = dioxin level in people's blood
  - X = exposure to contaminated soil along the river, age, sex, BMI, eating fish, water-related activity, occupation……

# Examples of Longitudinal Studies

## What is a longitudinal study?

A longitudinal study is a cohort study (follow-up study), in which repeated measurements are taken over time for each individual.

## Example 1. ( Framingham Study: Continuous outcome)

- In the Framingham heart study, each of 2634 study participants was examined every two years for 10 years for cholesterol level, BMI, smoking etc.
- The outcome cholesterol is a continuous variable.

## For each subject:

| | 0 | 2 | 4 | 6 | 8 | 10 | (year) |
|---|---|---|---|---|---|---|---|
| cholesterol | X | X | X | X | X | X | |
| BMI | X | X | X | X | X | X | |
| smoking | X | | | | | | |
| age | X | | | | | | |

## Study objectives:

1. How does cholesterol level change over time?

2. How is cholesterol level associated with BMI, smoking and age?

## Remarks:

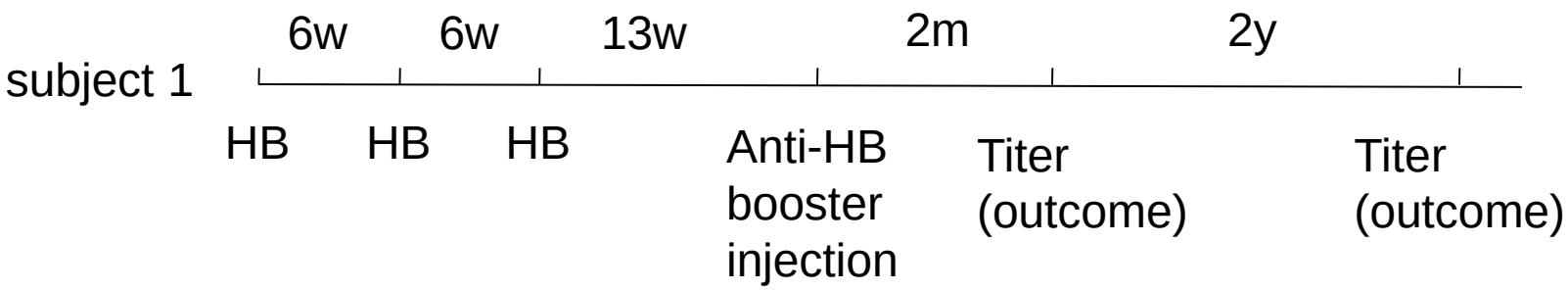
1. In a cross-sectional study, cholesterol level, BMI, smoking, age are all measured at a single time point. (e.g. time=0)

2. In this example, each subject has the same number of repeated measures at the same time points. However, in many longitudinal studies, observations may be obtained at irregular time points that vary between subjects and the number of observations may vary from one subject to another as well.

## Example 2. ( Hep-B vaccine study: Continuous outcome)

- 118 infants in Senegal.
- Each infant received 3 HB vaccine, 6 weeks apart.
- 13 weeks later, their degree of immunity to HB was measured by the titer to anti-hepatitis B virus surface antigen (Anti-HB) and a booster injection was then administrated.
- They were followed periodically from 2 to 67 months (about 6 years)
- Outcomes (Anti-HB) are continuous variables.

| | # of measurements over time | | | | | Total |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | |
| **# of subjects** | 43 | 33 | 24 | 12 | 6 | 118 |



- Study objectives:

1. What is the time course of anti-HBS conferred by vaccination?

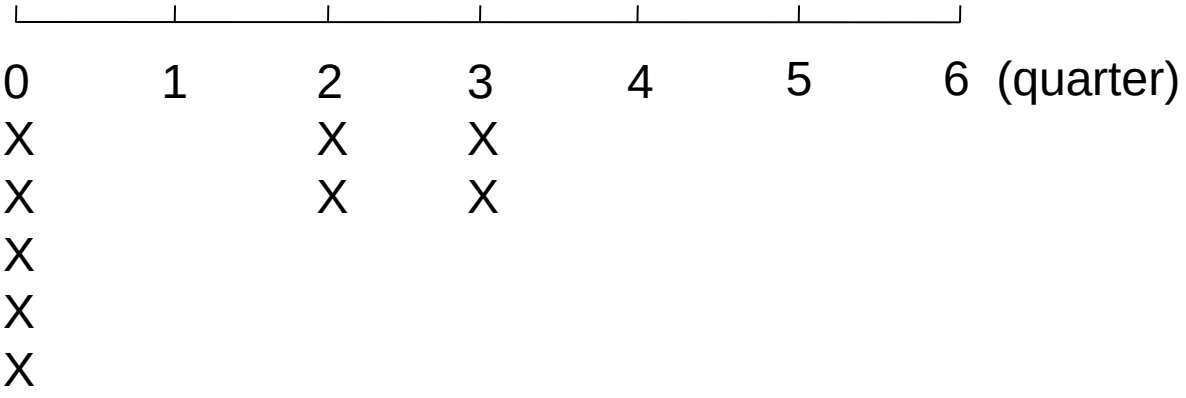
2. What is the effect of the pre-immunization anti-HBS titer on the anti-HBS levels after booster injection?

## Example 3. ( Respiratory infection study: Binary Outcome)

- 275 Indonesian pre-school children.
- Each was examined up to 6 consecutive quarters for the presence of respiratory infection (yes/no).
- Outcomes is binary.
- Covariates:
  - Age, sex, height for age, xerophthalmia (yes/no) (Vitamin A deficiency)

## Data:

(subject 1)




## Study objectives:

- Is the risk of respiratory infection related to vitamin A deficiency adjusted for age, sex, height effects?

# Comparison of Data Structures in cross-sectional and longitudinal studies

| | Cross-sectional Study | | Longitudinal Study | |
|---|---|---|---|---|
| | Subject | Data | Subject | Data |
| X=BMI | 1 | $X_1$ | 1 | $X_{11}\ X_{12}\ \ldots\ X_{15}$ |
| Y=CHOL | | $Y_1$ | | $Y_{11}\ Y_{12}\ \ldots\ Y_{15}$ |
| | 2 | $X_2$ | 2 | $X_{21}\ X_{22}\ \ldots\ X_{25}$ |
| | | $Y_2$ | | $Y_{21}\ Y_{22}\ \ldots\ Y_{25}$ |
| | …… | | …… | |

Note:

- Classical linear/logistic models for cross-sectional data has a single measure of X and Y for each subject, while longitudinal studies have multiple measures of X and Y for each subject.

- We here only consider a single independent variable.

# Why longitudinal studies?

1. A longitudinal study allows us to study the change of the outcome variable over time.

2. A longitudinal study is more powerful to detect an association of interest compared to a cross-sectional study.
   - Sample size
   - Each subject serves as his/her own control

3. A longitudinal study allows us to estimate the between-subject variation and the within-subject variation.

# Challenges in Analyzing Longitudinal Data

Recall:

- In classical linear/logistic regression, the key assumption is the independence between observations.
- For example,

$$Y_i = \beta_0 + \beta_1 X_{1\mathrm{i}} + \beta_2 X_{2\mathrm{i}} + \ldots + \beta_p X_{pi} + error$$

  - Y=SBP, $X_1$=age, $X_2$=sex…..
  - 1st subject's SBP does not depend on 2nd subject's SBP

However,

  - In a longitudinal study, the observations over time from the same subject are likely to be correlated.

- The observations from the same subject are likely to be similar in a longitudinal study.

- For example,

| | | time | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| SBP | Subject 1 | 150 | 142 | 157 |
| | Subject 2 | 100 | 105 | 97 |

  - Subject 1 is independent of subject 2.
  - Observations for the same subject are correlated.

- Classical linear/logistic regression analysis is INVALID, as standard error estimates of regression coefficients are often too small.

- A valid statistical inference must take the within-subject correlation into account.

# Example of Hierarchical Data

- Detroit Asthma Invention Study
  - About 360 children with asthma from 12 middle schools in Detroit (6 intervention (asthma mangement) schools and 6 control schools)
  - Each child was followed for 3 time points (0, 6, 12 months).
  - Data:

| School | child | time |
|---|---|---|
| 1 | 1 | x x x |
| | 2 | x x |
| …… | …… | |
| | 34 | x x . |
| 12 | …… | …… |
| | 360 | x x |

  - Objectives:
    - To study the intervention effects on asthma severity improvement.
    - To estimate between-school and between-child variabilities.

# Example of Spatial Data : disease mapping

- Scottish Lip Cancer Data ( Clayton and Kaldor 1987 Biometrics)
  - Data:
    - Observed and expected number of lip cancer cases in 56 counties of Scotland (SMR=O/E).
  - Covariates:
    - Percentage of the work force employed in agriculture, fishing, or forestry.
    - Measure of exposure to sunlight.

– **Objectives:**

- How is the standardized mortality rate (SMR) related to the degree of the exposure to sunlight?

- To produce a map to show “smooth” regional variation in lip cancer incidence and to avoid presenting unstable rates for small countries.

– **Challenge:**

- Spatial correlation.

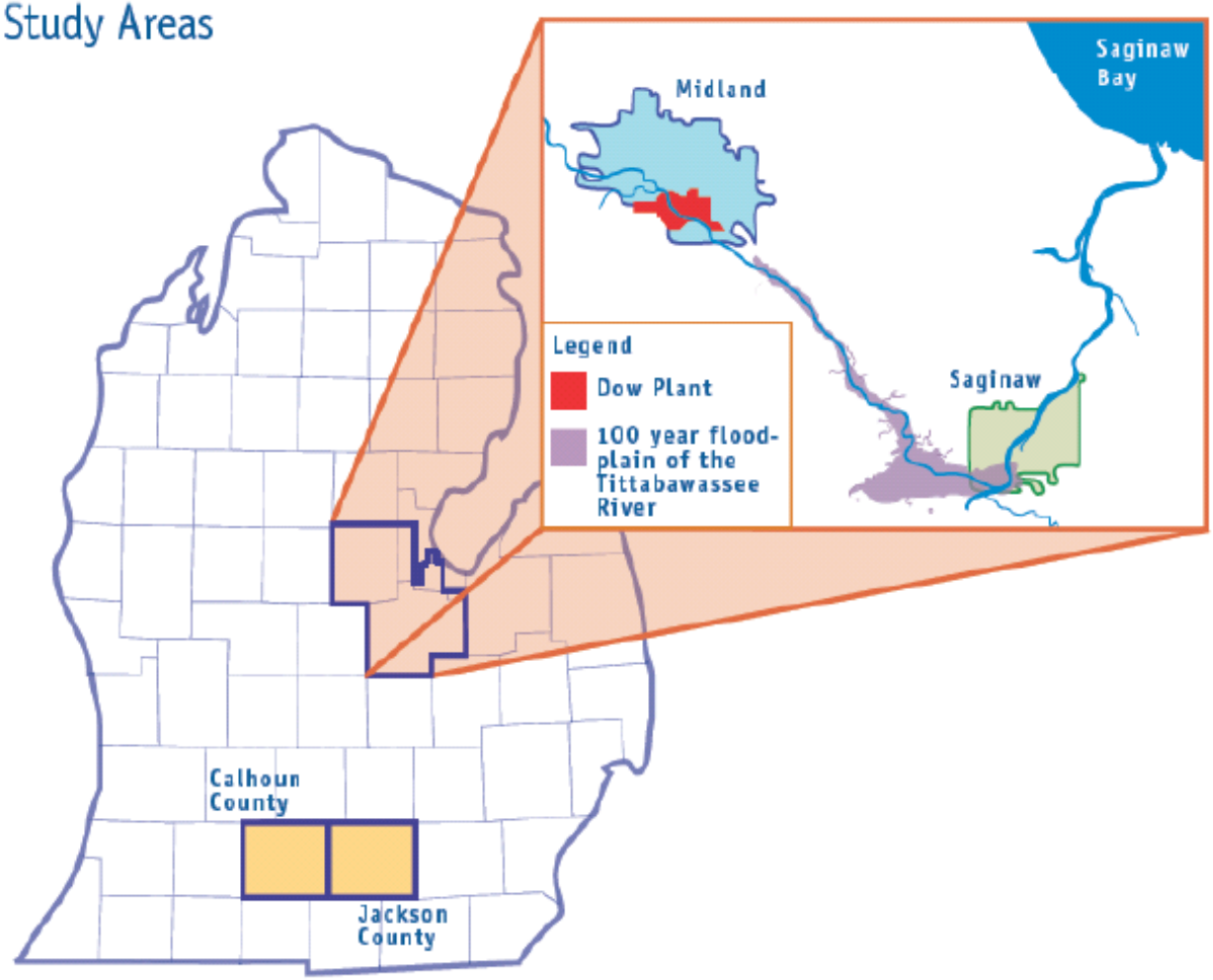
- # of cancer cases in adjacent counties may be correlated, since the subjects may share similar environment.

# Example of Spatial Data: Ecological Study

The University of Michigan Dioxin Exposure Study

- Elevated levels of dioxins have been found in the soil of the Tittabawassee River flood plain and nearby areas, as well as in residents' bodies.
- Objective: To evaluate factors that potentially caused high levels of dioxins in residents' bodies.
- Y = dioxin level in people's blood
- X = exposure to contaminated soil along the river, age, sex, BMI, eating fish, water-related activity, occupation……

# Study Areas


Saginaw Bay
Midland
Legend
Dow Plant
100 year flood-plain of the Tittabawassee River
Saginaw
Calhoun County
Jackson County

# Number of Participants in U-M Dioxin Exposure Study

| | Floodplain | Near Floodplain | Midland Plume | Other Midland/ Saginaw | Jackson/ Calhoun | Total Across All Areas |
|---|---|---|---|---|---|---|
| Interviews | 314 | 276 | 66 | 309 | 359 | 1324 |
| Blood Samples | 243 | 205 | 43 | 204 | 251 | 946 |
| Household Dust Samples | 205 | 161 | 32 | 168 | 198 | 764 |
| Soil Samples | 203 | 164 | 32 | 173 | 194 | 766 |
| Interviews, blood, dust and soil | 195 | 156 | 30 | 167 | 183 | 731 |

# Models for Correlated Data

- Generalized Estimating Equations (GEEs)
  - Liang and Zeger (1986, *Biometrika*)
  - Model: $$g(\mu_{ij}) = x_{ij}^{T}\beta$$

    Where g(.) is a link function.
  - Features:
    - Marginal Model (Population-Average Model), i.e. modeling the population mean.
    - Extension of generalized linear models to multivariate setting.

- The regression parameters are consistent and asymptotically normal given the mean model is correctly specified, no matter whether the correlation structure is correctly or incorrectly specified.

- Inference is robust to misspecification of the correlation structure and the sandwich estimator is used to correct for misspecification of the correlation structure.

- No likelihood. Likelihood-based inference (e.g., LR test) is difficult.

# Models for Correlated Data

- Transition model
  - Model:

$$g(\mu_{ij}) = x_{ij}^{T}\beta + \gamma_1 y_{i,j-1} + \ldots + \gamma_k y_{i,j-k}$$

  - Features:
    - Use the history to predict the future.
    - The conditional distribution of each response is expressed as an explicit function of the past responses and covariates.

# Models for Correlated Data

- (Generalized) Linear Mixed Model (GLMMs)
  - Laird and Ware (LMMs, 1982, *Biometrics*)
  - Breslow and Clayton (GLMMs, 1993, *JASA*)
  - Model:

$$g(\mu_{ij}) = x_{ij}^T \beta + z_{ij}^T b_i$$

  - More generally,

$$g(\mu_i) = x_i^T \beta + z_i^T b$$

where

$$b \sim N(0, D(\theta))$$

– Features:

- Random effect model (Subject-specific model).
- Model correlation explicitly using random effects in the modeling stage.
- Likelihood inference is straightforward in principle.
- Difficulty: likelihood is

$$L(\beta,\theta) = \int L(y \mid b) L(b) db$$

  Often involves intractable numerical integration.

- Convenient to use for subject-specific predictions.
- Inference requires a correct specification of the correlation structure.