Some key statistical ideas for 655 and 656
Longitudinal data

- Each subject gives rise to a vector of measurements representing the same response measured at a sequence of observation times
- Repeated responses over time on independent units (persons or cluster)
Basic issues and exploratory analyses
- Definition and examples of LDA
- Approaches to LDA
- Exploring correlation

Statistical methods for continuous measurements
- General Linear Model with correlated data
  - Weighted Least Squares estimation
  - Maximum Likelihood estimation
  - Parametric models for covariance structure

Generalized linear models for continuous/discrete responses
- Marginal Models
  - Log Linear Model and Poisson Model for count responses
  - Logistic model for binary responses
  - GEE estimation methods
  - Estimation techniques
- Random Effects Models (Multi-level models)
- Transition Models
Why special methods for LDA?

• Repeated observations $y_{i1}, y_{i2}, \ldots, y_{in}$ are likely to be correlated, so assumption of independence is violated

• What if we used standard regression methods anyway (ignore correlation)?
  – Correlation may be of scientific focus
  – Incorrect inference
  – Inefficient estimates of the association between the predictors $x$ and the outcome $y$
Characteristics of a LDA data set (an example of a clustered data set)

- There are repeated observations on each experimental unit
- Units (clusters) can be assumed independent of one another
- Multiple responses within each unit (cluster) are likely to be correlated
- The objectives can be formulated as regression problems whose purpose is to describe the dependence of the response on explanatory variables
- The choice of the statistical model must depend on the type of the outcome variable
Why LDA?

(A special case of a multilevel data set)

- Repeated measures made on the same subject will be correlated.
  - John’s cholesterol in 1989 will be related to John’s cholesterol in 1999.

- For subject $i$ at time $j$:
  - $Y_{ij} = X\beta + \varepsilon_{ij}$ with $\text{Corr}(\varepsilon_{ij}, \varepsilon_{ik}) \neq 0$

- To have correct inferences, must account for
Generalized Linear Models for Longitudinal Data

- Generalized Linear Models: A Review
- The Logistic Regression Model
  - Marginal model
  - Random effects model

Regression parameters have a different interpretation
GLM Examples

- Linear regression
  \[ \mu_i = X_i \beta; \quad g(\mu_i) = \mu_i \]
  \[ Y_i \sim N(\mu_i, \sigma^2) \]

- Logistic regression
  \[ \log \left( \frac{\mu_i}{1 - \mu_i} \right) = X_i \beta; \quad g(\mu_i) = \log \left( \frac{\mu_i}{1 - \mu_i} \right) \]
  \[ Y_i \sim \text{Bernoulli}(\mu_i) \]

- Poisson regression
  \[ \log \mu_i = X_i \beta; \quad g(\mu_i) = \log \mu_i \]
  \[ Y_i \sim \text{Poisson}(\mu_i) \]
Note for GLMs

• \( \text{var}(Y_i) \) may be a function of \( \mu_i \)
  
  - Logistic: \( \text{var}(Y_i) = \mu_i(1- \mu_i) \)
  
  - Poisson: \( \text{var}(Y_i) = \mu_I \)
1. Marginal Logistic Regression Model
   (use GEE for parameter estimation)

• **Goal**: To assess the dependence of respiratory infection on vitamin A status in the Indonesian Children’s Health Study

\[
x_{ij} = 1 \quad \text{if child } i \text{ is vitamin A deficient at visit } j
\]

\[
y_{ij} = 1 \quad \text{child } i \text{ has respiratory infection at visit } j
\]

\[
\mu_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)
\]

\[
\logit \mu_{ij} = \beta_0 + \beta_1 x_{ij}
\]

\[
P(Y_{ij} = 1) = \frac{\exp(\beta_0 + \beta_1 x_{ij})}{1 + \exp(\beta_0 + \beta_1 x_{ij})}
\]

\[
\text{var}(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})
\]

\[
\text{corr}(Y_{ij}, Y_{ik}) = \alpha
\]
Parameter Interpretation

- \( \frac{\exp(\beta_0)}{1+\exp(\beta_0)} = P(Y_{ij} = 1 \mid x_{ij} = 0) \) probability of infected children among the subpopulation that is not vitamin A deficient
- \( \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} = P(Y_{ij} = 1 \mid x_{ij} = 1) \) probability of infected children among the subpopulation that is vitamin A deficient
- \( e^{\beta_0} = \frac{P(Y_{ij}=1|x_{ij}=0)}{P_{T}(Y_{ij}=0|x_{ij}=0)} \) ratio (odds) of the probabilities of infected to uninfected children among the subpopulation that is not vitamin A deficient
- \( e^{\beta_0+\beta_1} = \frac{P(Y_{ij}=1|x_{ij}=1)}{P_{T}(Y_{ij}=0|x_{ij}=1)} \) ratio (odds) of the probabilities of infected to uninfected children among the subpopulation that is vitamin A deficient
- \( e^{\beta_1} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \text{odds of infection among vitamin A deficient children divided by the odds among children replete with vitamin A (odds ratio)} \)
- \( \beta_1 = \log \text{odds ratio} \)
Correlation between binary outcomes within the cluster

Two options:
1. Specify pairwise correlations
   \[ \text{corr}(Y_{ij}, Y_{ik}) = \alpha \]

2. Model association among binary data using the odds ratio
   \[ \text{OR}(Y_{ij}, Y_{ik}) = \frac{P(Y_{ij}=1, Y_{ik}=1)P(Y_{ij}=0, Y_{ik}=0)}{P(Y_{ij}=1, Y_{ik}=0)P(Y_{ij}=0, Y_{ik}=1)} \]

Which is better? Option 2.
2. Logistic model with random effects

Assume:

- The propensity for respiratory infections varies across children, reflecting their different genetic predispositions and unmeasured influences of environmental factors.
- Each child has his/her own propensity for respiratory disease $\beta^*_0 + U_i$, but that the effect of vitamin A deficiency ($\beta_1^*$) on this probability is the same for every child, i.e.

$$\logit P(Y_{ij} = 1 \mid U_i) = (\beta^*_0 + U_i) + \beta^*_1 x_{ij}$$

$$U_i \sim N(0, \sigma^2)$$

- Given $i$, we further assume that the repeated observations for the $i$th child are independent of one another.
Logistic model with random effects (cont’d)

\[ \beta_0^* = \log \text{odds of respiratory infection for a “typical” child (with random effect } U_i = 0) \]
Logistic model with random effects (cont’d)

\[ \exp(\beta^*_1) = \text{odds of infection for a child with random effect } U_i \text{ when he/she is vitamin A deficient relative to when the same child is not vitamin A deficient} \]

\[
\exp(\beta^*_1) = \frac{\exp(\beta^*_0 + \beta^*_1)}{\exp(\beta^*_0 + U_i)} = \frac{P(Y_{ij}=1|U_i,x_{ij}=1)/P(Y_{ij}=0|U_i,x_{ij}=1)}{P(Y_{ij}=1|U_i,x_{ij}=0)/P(Y_{ij}=0|U_i,x_{ij}=0)}
\]

\[ \nu^2 = \text{degree of heterogeneity across the children in the propensity of disease, not attributable to } x \]

\[
\exp(\beta_1) = \frac{P(Y_{ij}=1|x_{ij}=1)/P(Y_{ij}=0|x_{ij}=1)}{P(Y_{ij}=1|x_{ij}=0)/P(Y_{ij}=0|x_{ij}=0)}
\]
Random effects (RE) model: Basic ideas

- There is natural heterogeneity across individuals in their regression coefficients, and this heterogeneity can be explained by a probability distribution.
- RE models are most useful when the objective is to make inference about individuals rather than the population average.
- $\beta^*_1$ represents the effects of the explanatory variables on an individual child’s chance of infection.
  - ...this is in contrast with the marginal model coefficients, which describe the effect of explanatory variables on the population average.
Parameter Interpretation of a Logistic Regression Model with **Random Effects**

- $\logit P(Y_{ij} = 1 \mid U_i, x_{ij} = 1) = \beta_0^* + U_i + \beta_1^*$
- $\logit P(Y_{ij} = 1 \mid U_i, x_{ij} = 0) = \beta_0^* + U_i$
- $\text{Od}(U_i, x_{ij} = 1) = \frac{P(Y_{ij}=1|U_i,x_{ij}=1)}{P(Y_{ij}=0|U_i,x_{ij}=1)} = \exp(\beta_0^* + U_i + \beta_1^*)$
- $\text{Od}(U_i, x_{ij} = 0) = \frac{P(Y_{ij}=1|U_i,x_{ij}=0)}{P(Y_{ij}=0|U_i,x_{ij}=0)} = \exp(\beta_0^* + U_i)$
- $\text{Od}(U_i, x_{ij} = 1) = e^{\beta_1^*} \times \text{Od}(U_i, x_{ij} = 0)$

**Note:** The odds of respiratory infection for a hypothetical child with random effect $U_i$ and **with** vitamin A deficiency, are equal to $e^{\beta_1^*}$ times the odds of respiratory infection for the same hypothetical child with random effect $U_i$ **without** vitamin A deficiency.
Parameter Interpretation of a Logistic Regression Model with Random Effects (cont’d)

Compare the individual odds from the previous slide:

\[ Od(U_i, x_{ij} = 1) = \frac{P(Y_{ij} = 1|U_i, x_{ij} = 1)}{P(Y_{ij} = 0|U_i, x_{ij} = 1)} = \exp(\beta^*_0 + U_i + \beta^*_1) \]
\[ Od(U_i, x_{ij} = 0) = \frac{P(Y_{ij} = 1|U_i, x_{ij} = 0)}{P(Y_{ij} = 0|U_i, x_{ij} = 0)} = \exp(\beta^*_0 + U_i) \]
\[ Od(U_i, x_{ij} = 1) = e^{\beta_1} \times Od(U_i, x_{ij} = 0) \]

...with the population average odds below:

\[ Od(x_{ij} = 1) = \frac{P(Y_{ij} = 1|x_{ij} = 1)}{P(Y_{ij} = 0|x_{ij} = 1)} = \exp(\beta_0 + \beta_1) \]
\[ Od(x_{ij} = 0) = \frac{P(Y_{ij} = 1|x_{ij} = 0)}{P(Y_{ij} = 0|x_{ij} = 0)} = \exp(\beta_0) \]
\[ Od(x_{ij} = 1) = e^{\beta_1} \times Od(x_{ij} = 0) \]
In summary

1. Marginal model:

\[
\text{logit } P(Y_{ij} = 1) = \beta_0 + \beta_1 x_{ij}
\]

\(\beta_1\) describes the effect of explanatory variables on the chance of infection in the *entire population*.

2. Random effects model

\[
\text{logit } P(Y_{ij} = 1 \mid U_i) = \beta^*_0 + U_i + \beta^*_1 x_{ij}
\]

\(\beta^*_1\) describes the effect of the explanatory variables on an *individual* chance of infection.
Contrasting Approaches

- In **linear** models, the interpretation of $\beta$ is essentially independent of the correlation structure.
- In **non-linear** models for discrete data, such as logistic regression, different assumptions about the source of correlation can lead to regression coefficients with distinct interpretations.
- Two examples:
  - Infant growth
  - Respiratory disease data
In summary

1. Marginal model:

   \[ E[Y_{ij}] = \beta_0 + \beta_1 x_{ij} \]

   \( \beta_1 \) describes the change in the average response for a unit change in \( x_{ij} \) for the entire population

2. Random effects model

   \[ E(Y_{ij} \mid U_i) = \beta_0^* + U_i + \beta_1^* x_{ij} \]

   \[ E[Y_{ij}] = E[E(Y_{ij} \mid U_i)] = \beta_0^* + E[U_i] + \beta_1^* x_{ij} = \beta_0^* + \beta_1^* x_{ij} \]

   \( \beta_1^* \) describes the change in the average response for a unit change in \( x_{ij} \) for a particular subject, and for the entire population
Marginal Model vs. Random Effects

- The interpretation of the model parameters is different in marginal and random effects models for binary outcomes parameters
  - Marginal: ratio of population odds
  - Random Effects: ratio of individuals’ odds
- Marginal parameter values are small in absolute values than their random effects analogues

\[ | \beta_k | \leq | \beta^*_k | \]
Key concepts

- What is a longitudinal data set?
- What is a GLM?
- What is a marginal (population average) model?
- What is a conditional (random effects) model?
- Parameter interpretation under a marginal and a conditional model