BIOS 667

Regression Models for Longitudinal and Correlated Responses

Bahjat F. Qaqish
Department of Biostatistics
CB 7420, McGavran Greenberg Hall
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7420
email address: qaqish@bios.unc.edu

Outline

- Objectives of this review.
- Examples
- Are special methods needed for correlated responses?
- Analysis focus - mean vs correlation
- Regression models for independent responses: Linear, logistic, loglinear
- Generalized Linear Models for independent responses
- Three basic extensions for dependent (correlated) responses:
  - Marginal (Population-Average) models
  - Random-effects (Subject-Specific, Mixed) models
  - Conditional models

Objectives of this review

- The main objective is to introduce the major types of models used in the analysis of dependent responses.
- Emphasis is on interpretation of model parameters. The regression parameters β's in differnet model types have different interpretations.
- In study design and analysis, one of the key issues is choosing the type of model that most directly addresses the research questions.
- Mathematical and computational details are not addressed.

Example: The 6-City Study

- Harvard Study of Air Pollution and Health (Ware et al., 1984, Am. Rev. Resp. Dis).
- Data on 537 children from Stubenville, Ohio, examined from age 7 to age 10.
- Outcome: Respiratory infection in the year prior to the exam reported by the mother.
- Data:

```
Y7  Y8  Y9  Y10 MS  Count
0   0   0   0   1   237
0   0   0   1   1    10
0   0   1   0   1    15
0   0   1   1   1     4
0   1   0   0   1    16
0   1   0   1   1     2
0   1   1   0   1     7
...
1   1   1   0   2     4
1   1   1   1   2     7
```

Example: Epileptic Seizures

- A clinical trial involving 59 epileptics (Thall and Vail, 1990 Biometrics).
- Number of epileptic seizures during a baseline 8 week period was recorded.
- Randomization to progabide or placebo adjuvant treatment.
- Epileptic seizure count in four consecutive two-week intervals.
- Data:

```
 Y1  Y2  Y3  Y4  Treatment Baseline Age
  5   3   3   3     0          11     31
  3   5   3   3     0          11     30
  2   4   0   5     0           6     25
...
102  65  72  63     1         151     22
...
  0   0   0   0     1          13     36
  1   4   3   2     1          12     37
...
```

Example: ignoring correlation

- Suppose $Y_1, Y_2, \cdots, Y_n$ are correlated normally distributed observations with mean μ and variance σ^2, with all pairwise correlations equal to ρ. We want to construct a 95% confidence interval for the mean μ. For simplicity, assume that σ^2 and ρ are known.
- An interval with 95% coverage is

$$\bar{Y} \pm 1.96 \left[\frac{\sigma^2}{n} \{1 + (n-1)\rho\} \right]^{1/2}.$$

- If we wrongly assumed independence we would compute the interval

$$\bar{Y} \pm 1.96 \left[\frac{\sigma^2}{n} \right]^{1/2}.$$

- Actual confidence (coverage probability) % of the "wrong" interval:

ρ	0	0.2	0.4	0.6	1
n					
4	95	88	81	76	67
8	95	79	68	61	51
12	95	73	60	52	43

- Actual confidence can be much lower than the nominal level if correlation is ignored.
- For hypothesis tests, actual type-I error (α) can be much higher than the nominal level if correlation is ignored.

Example: ignoring correlation

- In one study, the number of observations per subject n_i ranged from 1 to 20. The within-subject correlation is ρ.
- The mean of each subject's observations is computed. We wish to combine data from all subjects to estimate the population mean. How?
- Weight = 1: Giving all subjects the same weight doesn't seem right. A subject with 10 observations obviously provides more information than a subject with 1 observation. Subjects with more data ought to get more weight. How much more?
- Weight = n_i: Giving each subject weight proportional to the number of observations doesn't seem to be a good idea. It would mean that the subject with 20 observations is providing as much information as 20 subjects with 1 observation each.
- Weight = 1/variance: This is the "optimal" (gives the most precise estimate) weight = $\frac{n_i}{1+(n_i-1)\rho}$.
- Relative precision (efficiency) % of different weights relative to the optimal weights:

n_i	1,20	1,20	1-20	1-20
weight	1	n_i	1	n_i
ρ				
0.0	18	100	53	100
0.1	44	95	78	94
0.2	62	88	88	90
0.3	75	82	93	87
0.4	84	77	96	84
0.5	90	72	98	83
0.6	94	68	99	81
0.7	97	64	99	80
0.8	99	61	100	79
0.9	100	58	100	78
1.0	100	55	100	77

- Efficiency has an "effective sample size" interpretation; an estimator with 70% efficiency used with data from 100 subjects achieves the same variance as an estimator with 100% efficiency and data from 70 subjects.
- Summary: Special methods are needed for analysis of correlated responses:
  - To obtain valid statistical inferences (main reason).
  - To obtain more efficient statistical inferences; more precision. (less important, icing on the cake).

Analysis focus - mean vs correlation

- It is important at the outset to decide what the focus is.
- Focus on the mean structure: We would like to take correlation into account, although it is not the main focus of the analysis.
- Focus on the correlation structure: Careful modelling of correlation is required. Example: a study of the familial correlation of COPD. Risk of COPD (mean structure) is known to depend on smoking, age, etc. However, the main interest is modeling the sib-sib, sib-father and sib-mother correlations as a function of covariates.

Linear regression

- Response Y_i, covariates x_i on the i-th subject; $i = 1, \cdots, n$ independent subjects. Individual covariates $x_{i1}, \cdots, x_{ip}$. Usually $x_{i1} = 1$, the intercept.
- Focus is on the expected value, the mean, $E[Y_i] = \mu_i$.

$$\begin{aligned} \text{observed} &= \text{expectd} + \text{error} \\ Y_i &= \mu_i + (Y_i - \mu_i). \end{aligned}$$

- Linear regression, the mean

$$E[Y_i] = \mu_i$$

equals the systematic component

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

- The random error

$$\text{var}(Y_i - \mu_i) = \text{var}(Y_i) = \sigma^2$$

Constant variance. Normality assumed sometimes.

- Interpretation: β_j is the increase in the expected value associated with 1 unit increase in the j-th covariate, when **all other covariates are held constant**.

Logistic regression (for 0/1 responses)

- Focus is on the expected value, the mean, $E[Y_i] = P(Y_i = 1) = \mu_i$.
- The logit of the mean
$$\log \frac{\mu_i}{1-\mu_i} = \text{logit}(\mu_i)$$
equals the systematic component
$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$
- The random error
$$\text{var}(Y_i) = \mu_i(1-\mu_i)$$
Variance changes with the mean (not constant). Bernoulli distribution.
- Interpretation: β_j is the increase in the logit of the expected value associated with 1 unit increase in the j-th covariate, when **all other covariates are held constant**.

Loglinear models for counts

- Focus is on the expected value, the mean, $E[Y_i] = \mu_i$.
- The log of the mean
  $$\log(\mu_i)$$
  equals the systematic component
  $$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$
- The random random error
  $$\mathrm{var}(Y_i) = \mu_i$$
  or more generally
  $$\mathrm{var}(Y_i) = \sigma^2 \mu_i$$
  Variance proportional to the mean.
- Interpretation: β_j is the increase in the log of the expected value associated with 1 unit increase in the j-th covariate, when **all other covariates are held constant**.

Generalized Linear Models

- Y_i is a random variable,

$$E[Y_i] = \mu_i,$$

$$\text{var}(Y_i) = \sigma^2 h(\mu_i).$$

  Variance proportional to a known function of the mean. $h()$ is the **variance function**.

- The systematic component:

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

- The **link function** links the random component to the systematic component.

$$g(\mu_i) = \eta_i$$

- Interpretation: β_j is the increase in the expected value, on the scale of the link function, associated with 1 unit increase in the j-th covariate, when **all other covariates are held constant**.

Notation for Correlated Responses:

- Example: The 6-City Study. Data on 537 children from Stubenville, Ohio, examined from age 7 to age 10.
  Response: Respiratory infection in the year prior to the exam reported by the mother.
  Covariates: Mother smoking, age.

- Y_{ij} := Response (outcome) for the i-th child at the j-th time,
  $Y_{ij} = 1$ is if respiratory infection was reported, $Y_{ij} = 0$ otherwise.
  $i = 1, \cdots, 537; j = 1, 2, 3, 4.$

- x_{ij} := Covariates for the i-th child at the j-th time,
  Example:
  $x_{ij1} = 1$,
  x_{ij2} = mother smoking status (0=no, 1=yes),
  x_{ij3} = age (years).

-

Data:

ID	obs	outcome	intercept	MS	Age
i	j	y_{ij}	x_{ij1}	x_{ij2}	x_{ij3}
17	1	0	1	1	7
17	2	1	1	1	8
17	3	1	1	1	9
17	4	0	1	1	10
23	1	1	1	0	7
23	2	0	1	0	8
23	3	1	1	0	9
23	4	0	1	0	10

- One child = one cluster
  Number of clusters = $K = 537$.
  Cluster size = number of observations in a cluster, n_i = size of the i-th cluster,
  generally varies from cluster to cluster.
  In this example $n_i = 4$ for all clusters.

- A more complicated situation: Suppose Y_{ij} represents 3 different outcomes (e.g. respiratory infection, diarrheal disease and body weight) observed at the j-th occasion, $j = 1, \cdots, 4$, For 4 occasions there will be 12 outcomes per subject. The notation can be modified so that j goes from 1 to 12, and at each j there is one outcome.

Three basic types of models

- Generalized Linear Models for independent responses have three basic extensions for dependent (correlated) responses:
  - Marginal (Population-Average) models
  - Conditional models
  - Random-effects (Subject-Specific, Mixed) models
- The names reflect the interpretation of the regression parameters β.
- Here we are concerned with populations, not samples.
- References:
  Generalized Linear Models, P. McCullagh and J. A. Nelder. 2nd ed. Chapman & Hall, London, 1989.
  Analysis of Longitudinal Data, Diggle, Heagerty, Liang & Zeger. 2nd ed. Oxford University Press, Oxford, 2002.

Marginal (Population-Average) models

- The familiar regression models (linear, logistic, etc) for independent responses are marginal regression models. Marginal regression parameters for dependent (correlated) responses retain the same interpretation, because they are models for the **marginal** mean, population average, of the response.

- Interpretation of $P(Y_{i1} = 1; MS = 1, AGE = 7) = E[Y_{i1} = 1; MS = 1, AGE = 7] = 0.2$: Envisage an infinitely large population of children all age 7 and all have smoker mothers. The average value of Y_{i1} in that population, i.e. the population average, is 0.2. In that population 20% of the Y_{i1}'s are 1 and 80% are 0's. A child picked at random from that population has probability 20% of having $Y_{i1} = 1$ and 80% of having $Y_{i1} = 0$.

- In the 6-city study:

$$\begin{aligned} &\text{logit } P(\text{respiratory infection for child } i \text{ at time } j) \\ &= \text{logit } P(Y_{ij} = 1; MS, AGE) = \beta_1 + \beta_2 MS + \beta_3 AGE \\ &= \text{logit } P(Y_{ij} = 1; x_{ij}) = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3}. \end{aligned}$$

  In expanded form:

$$\begin{aligned} \text{logit } P(Y_{i1} = 1) &= \beta_1 + \beta_2 MS + 7\beta_3, \\ \text{logit } P(Y_{i2} = 1) &= \beta_1 + \beta_2 MS + 8\beta_3, \\ \text{logit } P(Y_{i3} = 1) &= \beta_1 + \beta_2 MS + 9\beta_3, \\ \text{logit } P(Y_{i4} = 1) &= \beta_1 + \beta_2 MS + 10\beta_3. \end{aligned}$$

- The model is for the population averages of Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}, each average done "separately".

- The margins do not determine the joint distribution. Example:

	$Y_{i2} = 0$	$Y_{i2} = 1$	
$Y_{i1} = 0$			0.8
$Y_{i1} = 1$			0.2
	0.7	0.3	1.0

- β_2 is the difference, on the log-odds scale, between the risk of respiratory infection in a population of children of a given age whose mothers smoke and a population of children of the same age whose mothers don't smoke.

- These models are relevant when the main focus of a study is studying effects of covariates (e.g. age, race, intervention, genetic markers, treatment) on the population mean.

- To do actual estimation from samples, further assumptions are required. However, that doesn't affect the interpretation of the β's.

- GEE is a particularly attractive method of estimation in these models.

Conditional models

- A simple generic form:
$$g(E[Y_{ij}|f(Y_{i(-j)}); x_{ij}] = x_{ij}^t\beta + \gamma f(Y_{i(-j)}).$$
- Example: Children with otitis media (middle-ear infection) randomized to antibiotic A (x=0) or B (x=1). Reference: Rosner (1989, JASA).
- Let Y_R, Y_L be the right and left ear responses after 2 weeks, respectively, coded so that 1 = improved, 0 = didn't.
- We can regress each ear's response on X and "adjust for" the other ear's response. e.g.
$$\text{logit } P(Y_R = 1) = \alpha + \beta x + \gamma Y_L,$$
$$\text{logit } P(Y_L = 1) = \alpha + \beta x + \gamma Y_R.$$
- The above notation is sloppy! We should have written
$$\text{logit } P(Y_R = 1|Y_L) = \alpha + \beta x + \gamma Y_L,$$
$$\text{logit } P(Y_L = 1|Y_R) = \alpha + \beta x + \gamma Y_R,$$
to make it explicit that the probability on the left is **conditional**.
- Rule: If responses appear on the right side of a regression equation, the mean or probability on the left side is necessarily conditional.
- The above is a **conditional** regression model because it conditions one outcome on another. i.e. It is a model for the conditional probability of recovery in one ear given the outcome in the other ear.
- Interpretation of β in the above model: Stratify the population by the left ear response, within each stratum $\exp(\beta)$ is the odds ratio relating right ear recovery to receiving treatment B vs A. Similarly, stratify the population by the right ear response, within each stratum $\exp(\beta)$ is the odds ratio relating left ear recovery to receiving treatment B vs A.
- $\exp(\beta)$ represents the B vs A effect on one ear given the outcome of the other ear.
- Is $\exp(\beta)$ a useful measure of treatment effect?
- What does $\exp(\beta)$ mean for a child who has only 1 ear infected? See M&N, the top half of page 235.

Conditional models

- Example: 6-city study.
- We can regress each each year's outcome on the previous year's outcome and mother smoking:

$$\text{logit } P(Y_{i2} = 1|Y_{i1}) = \alpha + \beta MS_i + \gamma Y_{i1},$$

$$\text{logit } P(Y_{i3} = 1|Y_{i1}, Y_{i2}) = \alpha + \beta MS_i + \gamma Y_{i2},$$

$$\text{logit } P(Y_{i4} = 1|Y_{i1}, Y_{i2}, Y_{i3}) = \alpha + \beta MS_i + \gamma Y_{i3},$$

- Y's appear on the right, so it is a **conditional** regression model.
- Transitional models, autoregressive, Markov chains.
- Interpretation of β in the above model: $\exp(\beta)$ is the odds ratio measuring the effect of mother smoking on respiratory infection in one year conditional on the preceding years' outcomes.
- Good for estimation of future risk for a certain child given what we know now about "this particular child".
- Useful in clinical settings.

Conditional models

- In a family study of covariates x and outcome Y. One family (cluster) has binary outcomes: 0 1 1 0 1. Define a new "covariate" = z, which for each member is the number of other family members affectd: 3 2 2 3 2. These new "covariates" can be included in a regression model along with other legitimate covariates x. The slopes for x represent the effect of x **conditional** on the number of other family members affected z.
- Loglinear regression models are conditional (see M&N, ch. 6). Example: For the otitis media example consider the loglinear regression model with model formula:

  $$x + Y_R + Y_L + Y_R * Y_L + Y_R * x + Y_L * x.$$

  The $Y_R * x$ parameter is the treatment effect on the right ear conditional on the left ear outcome. The $Y_L * x$ parameter is the treatment effect on the left ear conditional on the right ear outcome.
- Easy to fit (mostly).
- Parameter interpretation can be specific or sensitive to cluster size. e.g. A family of size 2 with 2 members affected is not similar to a family of size 7 with 2 members affected.
- Not reproducible. e.g. The otitis media example:

  $$\text{logit } P(Y_R = 1 | Y_L) = \alpha + \beta x + \gamma Y_L.$$

  What happens when the left year is not infected? In this case Y_L is not 0 and not "missing". It simply doesn't exist. Nevertheless, based on the conditional model, we can derive the "marginal" regression of Y_R on x:

  $$\text{logit } P(Y_R = 1) = \alpha^* + \beta^* x.$$

  However, note that $\alpha^* \neq \alpha$ and $\beta^* \neq \beta$ except in the (trivial) case of independence of Y_R and Y_L at each x. Furthermore, generally logit $P(Y_R = 1)$ will not be linear in x.

Random-effects (Subject-Specific, Mixed) models

- It is reasonable to think that differnet children under similar conditions (age, mother smoking) have different risks of respiratory infection. This could be due to natural biologic variation in susceptibility. This can be modelled by assuming that the i-th child has its own subject-specific risk, which depends upon observed covariates and an unobservable random variable, or "random effect", U_{i1}:

  $$\text{logit } P(Y_{ij} = 1|U_{i1}; MS, AGE) = U_{i1} + \beta_2 MS + \beta_3 AGE.$$

  The intercept is specific to the i-th subject and reflects **unobserved heterogeniety**.

- U_{i1} is assumed to be randomly distributed across subjects with mean β_1 and variance σ^2.

- $exp(\beta_2)$ = the odds ratio relating MS to RI keeping AGE and U_{i1} constant, i.e. odds ratio of disease at a given age comparing the same child had its mother smoked to itself had its mother not smoked. The contrast is "within-subject". It could also be between-subject, e.g. between subjects i and k provided their random effects are equal, i.e. provided $U_{i1} = U_{k1}$. Since the random-effects are unobservable, it is hard to imagine how we can find two subjects with the same value of the random effect.

- The parameters β are called the "fixed effects", while σ^2 is a "variance component".

- $exp(\beta_2)$ = the odds ratio relating MS to RI keeping AGE constant and conditional on the subject-specific random effect U_{i1}.

- $exp(\beta_2)$ = the ratio of subject-specific odds of disease relating MS to RI keeping AGE fixed.

- σ^2 = population heterogeniety in risk of RI = between-subject variance. e.g. $\sigma^2 = 4, \sigma = 2$. The odds ratio of a subject 1σ above the mean relative to a subject exactly at the mean is $\exp(\sigma) = \exp(2) = 7.4$. The odds ratio of a subject 1σ above the mean to a a subject 1σ below the mean is a $\exp(2\sigma) = \exp(4) = 54.6$.
  Note: In a normal distribution, about 2/3 of the population fall within $\pm 1\sigma$.

  A very small σ^2 indicates very little variation in subject-specific risk among children with the same MS and AGE.

- Large σ implies high within-subject correlation, and vice-versa.

- General formulation: $Y_{i1}, \cdots, Y_{in_i}$ are assumed conditionally independent given U_i.

  $$\begin{aligned} \nu_{ij} &:= E[Y_{ij}|U_i; X_i, Z_i, \beta]), \\ g(\nu_{ij}) &= x_{ij}^\top \beta + z_{ij}^\top U_i, \\ \text{var}(Y_{ij}|U_i) &= \phi V(\nu_{ij}), \end{aligned}$$

  where $g(\cdot)$ is the link function, $V(\cdot)$ is the variance function, U_i has mean zero and its distribution depends on parameters Γ. The covariates Z are a subset of X.

- The marginal picture in *linear* random-effects models: Suppose the random effects are centered, $E[U] = 0$, and start with a linear random-effects model

  $$E[Y|U; x, z, \beta] = x^\top \beta + z^\top U.$$

  The induced marginal model is also linear

  $$E[Y; x, z, \beta] = x^\top \beta$$

  So, in a linear random-effects model the fixed effects, β, have both a conditional and a marginal interpretation. A linear random-effects model implies a linear marginal model with the same fixed effects (slopes).

- The marginal picture in *nonlinear* random-effects models: Consider the model

  $$g(E[Y|U; x, z, \beta]) = x^\top \beta + z^\top U,$$

  where $g(\cdot)$ is a nonlinear link function and U has mean zero and its distribution depends on parameters Γ. Generally, the induced marginal model is not linear, but rather a complicated nonlinear function, $c(\cdot)$, of the covariates (x, z) and the parameters (β, Γ),

  $$g(E[Y; x, z, \beta, \Gamma]) = c(\beta, \Gamma, x, z) \neq x^\top \alpha.$$

- Prediction of U_i and estimation of $E[U_{i1}|\text{observed data}]$ is possible - Empirical Bayes.
- Random slopes: Possibly, different children respond to MS differently?

  $$\text{logit } P(Y_{ij} = 1|U_{i1}, U_{i2}; MS, AGE) = U_{i1} + U_{i2}MS + \beta_3 AGE$$

  Both the intercept and MS effect (slope) are random effects specific to the i-th subject.
- Identifiability problems.
- Difficult to fit (computationally intensive). See McCulloch (1997, JASA, p162-170) and references therein.