課程名稱: 長期追蹤資料分析 (Longitudinal data Analysis)

老師姓名: 沈仲維　　研究室:448

Email: weya071444@yahoo.com.tw

評分標準: 作業 30%、期中報告 30%、期末報告:40%

參考用書:
Longitudinal Data Analysis　Garrett M. Fitzmaurice 1962- 2008 (圖書館)
上課筆記與相關參考論文

基礎複習: Linear Models
程式: SAS\ R code

(1) Introduction to Longitudinal data Analysis
- (a) Introduction and Examples
- (b) Early origins of linear models for longitudinal data analysis

(2) Generalized Linear Models (GLM)
- (a) Binomial Data: Logistic regression
- (b) Count Data: Poisson Regression

(3) Generalized Estimating Equations (GEE)
- (a) Quasi-likelihood
- (b) First-order generalized estimating equations (GEE1)
- (c) Second-order generalized estimating equations (GEE2)

(4) Special issues for GEE
- (a) Ordinal data
- (b) Missing data

(5) Linear mixed-effects models (LMM)
- (a) Likelihood Inference for Linear Mixed Models
- (b) Confidence Intervals and Hypothesis Tests

Introduction to Longitudinal Data

- **Longitudinal data** consist of observations (i.e., measurements) **taken repeatedly through time** on a sample of experimental units (i.e., individuals, subjects).
- Longitudinal data are to be contrasted with **cross-sectional data**. Cross-sectional data contain measurements on a sample of subjects at **only one point in time**.
- **Repeated measures:** The terms "repeated measurements" or, more simply, "repeated measures" are sometimes used as rough synonyms for "longitudinal data", however, there are sometimes slight differences in meaning for these terms.
- Repeated measures are also multiple measurements on each of several individuals, but they **are not necessarily through time**. E.g., measurements of chemical concentration in the leaves of a plant taken at different locations (low, medium and high on the plant, say) can be regarded as repeated measures.

Some examples:

1. Longitudinal data – respiratory illness

The data are from a clinical trial of patients with respiratory illness, where 111 patients from two different clinics were **randomized** to receive either placebo or an active treatment. Patients were examined at baseline and at four visits during treatment. At each examination, respiratory status (categorized as 1 = good, 0 = poor) was determined.

(1) The recorded variables are:
Center (1,2), ID, Treatment (A=Active, P=Placebo), Gender (M=Male,F=Female), Age (in years at baseline), Baseline Response.

(2) The response variables are:
Visit 1 Response, Visit 2 Response, Visit 3 Response, Visit 4 Response.

	center	id	treat	sex	age	baseline	visit1	visit2	visit3	visit4
1	1	1	P	M	46	0	0	0	0	0
2	1	2	P	M	28	0	0	0	0	0
3	1	3	A	M	23	1	1	1	1	1
4	1	4	P	M	44	1	1	1	1	0
5	1	5	P	F	13	1	1	1	1	1
6	1	6	A	M	34	0	0	0	0	0
7	1	7	P	M	43	0	1	0	1	1
8	1	8	A	M	28	0	0	0	0	0

Table 1: Respiratory data for eight individuals. Measurements on the same individual tend to be alike.

Interest is in comparing the treatments, but also to include center, age, gender and baseline response in the model. From Table 1 it is clear, **that there is a dependency among the response measurements on the same person – measurements on the same person tend to be alike. This dependency must be accounted for in the modelling.**

核定機關：行政院主計處 96/03/15
核准文號：處普三字第 0960001524 號
有效期間：96 年 12 月 31 日止
辦理機關：行政院衛生署國民健康局

樣本編號：(訪員填寫)

A							
B							
C	鄉鎮區代號				序列號		

國民健康局 BUREAU OF HEALTH PROMOTION

民國 96 年中老年身心社會生活狀況長期追蹤(第六次)調查

個案姓名：_____　　個案之性別：□1 男　□2 女

相關參考論文

Shen, Chung-Wei; Chen, Yi-Hau* (2012) Model Selection for Generalized Estimating Equations Accommodating Drop-out Missingness. *Biometrics*, **68**: 1046–1054.

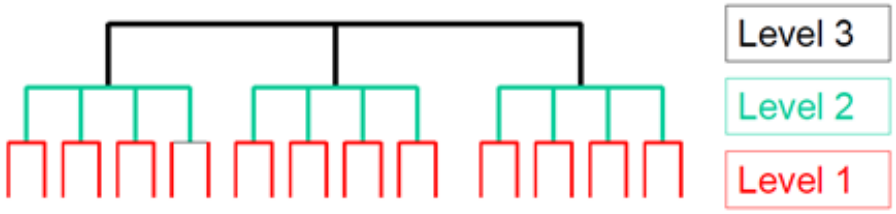
Shen, Chung-Wei; Chen Yi-Hau* (2013) Model Selection of Generalized Estimating Equations with Multiply Imputed Longitudinal Data. *Biometrical Journal*, **55**: 899-911.

Shen, Chung-Wei*, Chen, Yi-Hau (2015) Model Selection for Marginal regression analysis of longitudinal data with missing observations and covariate measurement error. *Biostatistics*, **16**: 740-753.

Shen Chung-Wei, Chen Yi-Hau* (2018) Joint model selection of marginal mean regression and correlation structure for longitudinal data with missing outcome and covariates. *Biometrical Journal*, **60:** 20-33.

2. Cluster Data: 群集型資料

2. Clustered/multilevel studies



最典型的教育研究的例子如下：若我們有興趣想了解學生間之學習成就的差異，**但是如果你有很多班級的話，每個班級的老師不同，這就衍生了問題：學生的學習成就可能是受到教師或班級影響(同一個班級內的學生是有相關性的)，所以我們想要解決缺乏獨立性 (lack of independence)的問題**。可以具體說說有什麼變數在班級裡面會影響學生成就? 這可說的太多了。比如說班級人數，男女生比例，貧窮學生比例。此外，老師當然也是重要的因素，比如說老師的經驗，老師的教育水準，老師的教學法等等。這樣一列下來，如果你想要列出一大堆變數來控制，似乎就顯得有點不切實際。更重要的是：你不可能控制所有的不同。從上面的例子來說，你就很容易可以看出來階層性關係。如果**學生是第一層（底層）的話，班級就是第二層（中層），學校就是第三層(上層)。**由於這個層次有階層性，所以在統計時就要列入考量，這也就是階層線性模式的最主要目的。如果你不熟悉 HLM 或是 multilevel linear models，也有可能在你的領域使用 mixed-effects models 、 random-effects models 或 random-coefficient regression models 等其它名詞。

2.2. Family data

Table 9.4. Response Counts of (Litter Size, Number Dead) for 58 Litters of Rats in a Low-Iron Teratology Study

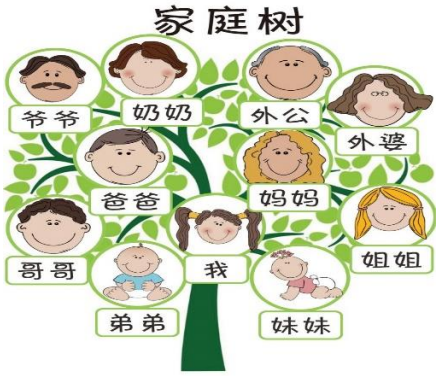
Group 1: untreated (low iron)
(10, 1) (11, 4) (12, 9) (4, 4) (10, 10) (11, 9) (9, 9) (11, 11) (10, 10) (10, 7) (12, 12)
(10, 9) (8, 8) (11, 9) (6, 4) (9, 7) (14, 14) (12, 7) (11, 9) (13, 8) (14, 5) (10, 10)
(12, 10) (13, 8) (10, 10) (14, 3) (13, 13) (4, 3) (8, 8) (13, 5) (12, 12)

Group 2: injections days 7 and 10
(10, 1) (3, 1) (13, 1) (12, 0) (14, 4) (9, 2) (13, 2) (16, 1) (11, 0) (4, 0) (1, 0) (12, 0)

Group 3: injections days 0 and 7
(8, 0) (11, 1) (14, 0) (14, 1) (11, 0)

Group 4: injections weekly
(3, 0) (13, 0) (9, 2) (17, 2) (15, 0) (2, 0) (14, 1) (8, 0) (6, 0) (17, 0)

Source: D. F. Moore and A. Tsiatis, *Biometrics*, **47**: 383–401, 1991.



相關參考論文

1. Chun Shu Chen, **Shen Chung-Wei*.** (2018) Model selection based on resampling approaches for cluster longitudinal data with missingness in outcomes. *Statistics in medicine*, **37:** 2982-2997.

2.3. A Dental Study of Periodontal Disease

(1) Informative Cluster Size

相關論文:

1. **Shen Chung-Wei**, Chen Yi-Hau*. (2018) Model selection for semiparametric marginal mean regression accounting for within-cluster subsampling variability and informative cluster size. *Biometrics*, **74**: 934-943.

2. Chien Li-Chu, Chang Li-Ying, **Shen Chung-Wei** .(2021) Model selection for correlated survival data with informative cluster sizes. (Submitted)

(2) Zero Inflated Models (零膨脹)

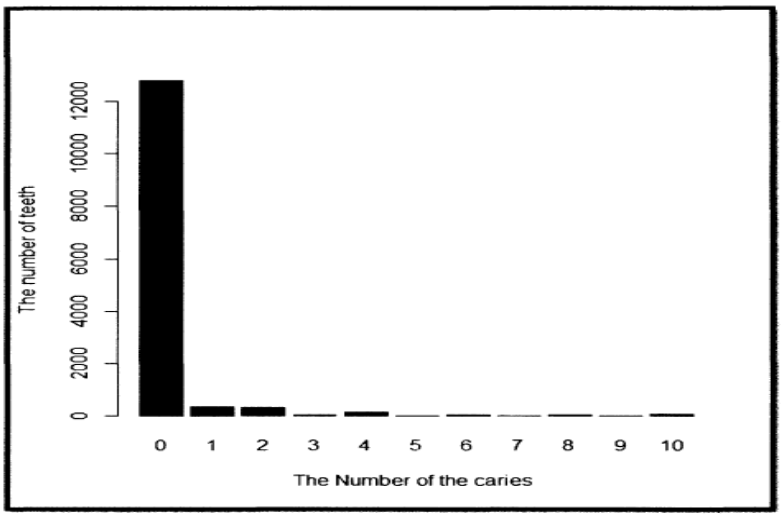


Figure 2. The frequencies of teeth with different number of caries indicate excessive zero counts for caries.

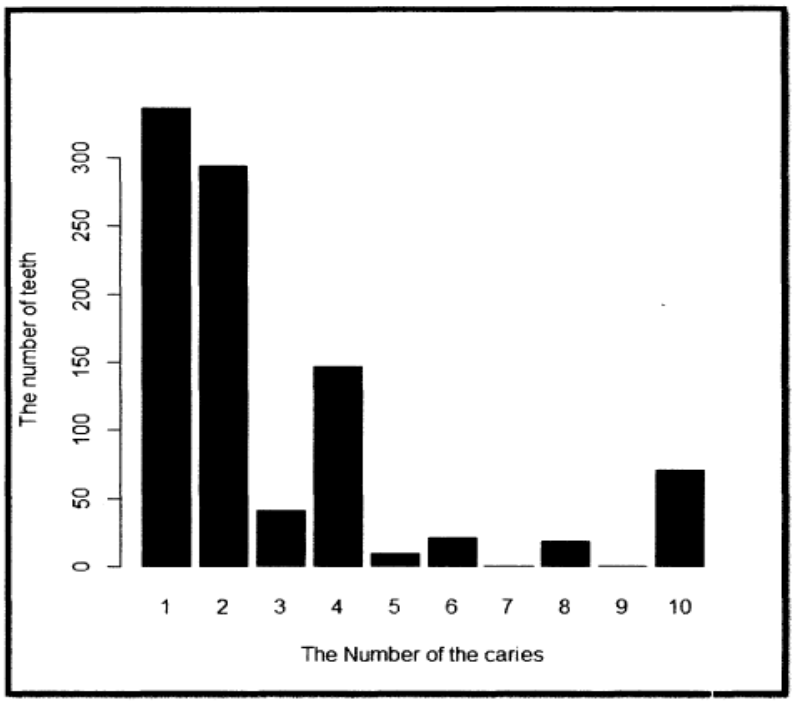
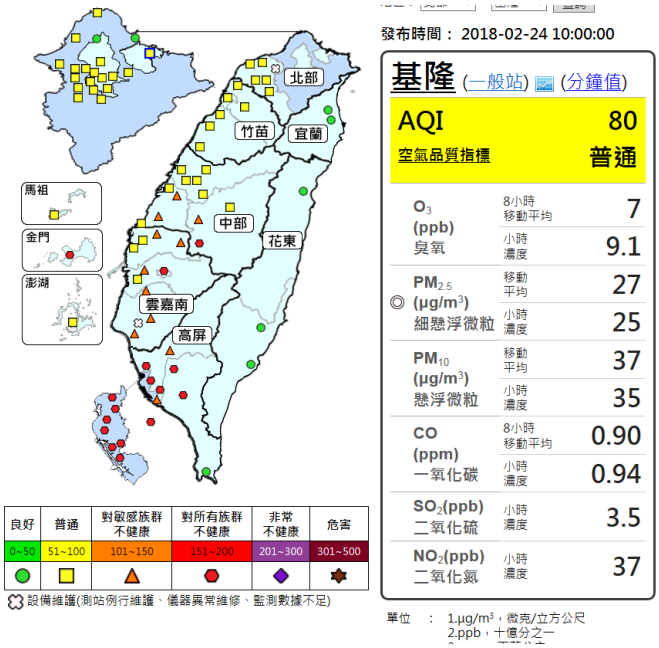


Figure 3. The frequency of teeth with different number of carries, excluding zero counts.

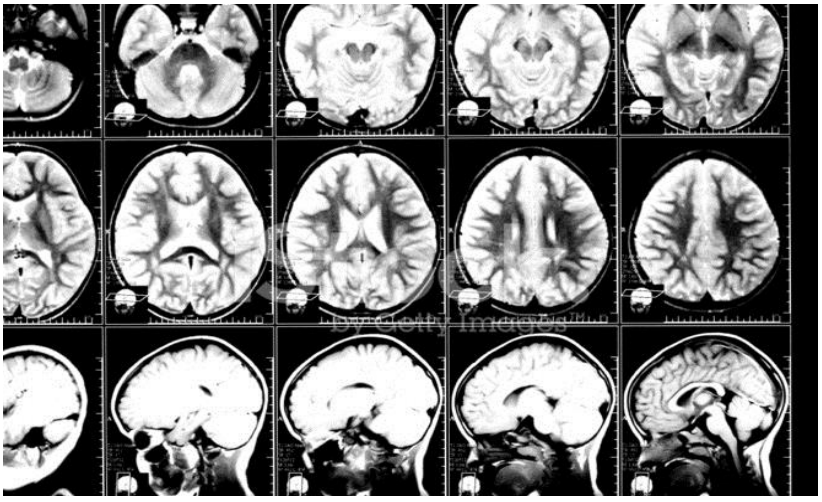
3. Spatial data　空間型資料



相關論文:

Shen Chung-Wei, Chen Yi-Hau*, Chen Chun Shu* (2021) Distribution-free regression model selection with a nested spatial correlation structure. *Spatial Statistics* **41**:100476.

4. Computed Tomography Data　電腦斷層掃描資料



Early origins of linear models for longitudinal data analysis

● Time Plots of Trends

A plot constructed this way displays how an outcome measurement changes over time, thereby providing information about trajectories or trends in the response.

(1) One describes **the pattern of change over time for individuals, displaying development in the response measurement for each subject.** The resulting longitudinal trajectories or curvatures of individuals give rise to a time plot of individual-based transitions, generally referred to as *intraindividual growth patterns*. Deviations in various intraindividual growth curves graphically display between-subjects variability in the response measurements. To compare general patterns of subject-specific growth across two or more population groups, the intraindividual time plots can be created separately by stratifying a discrete covariate, such as treatment, age group, gender, or race/ethnicity.

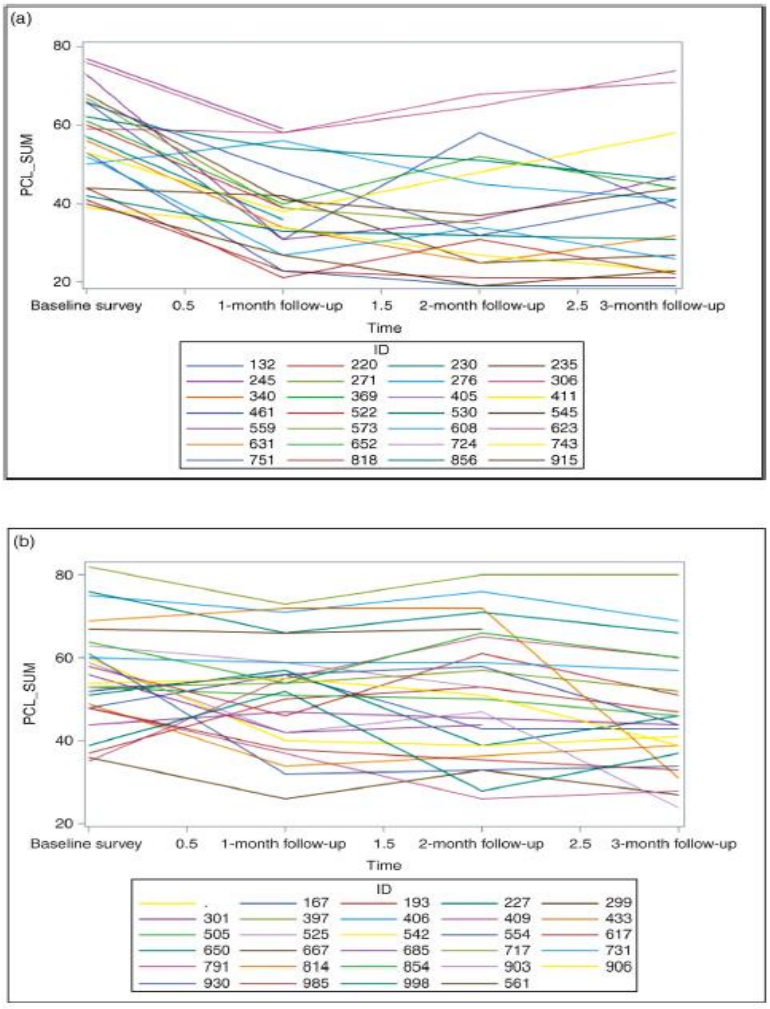


FIGURE 2.1

(a) Subject-specific time plot on PCL score for treatment group. (b) Subject-specific time plot on PCL score for control group.

(2) The second time plots perspective is the description of time trends for a population. In this procedure, the researcher calculates the average of the response measurements at each of the predesigned time points and then presents the mean scores in a time plot. The resulting plot for the sequence of means over time describes the pattern of change in the mean response score for the entire population of interest. One can also compute the **standard errors and corresponding confidence intervals,** and then display these supplementary statistics simultaneously in the time plot. This population trend approach is useful in disciplines where the **primary interest of research is in the entire population or a population subgroup, rather than in particular individuals**

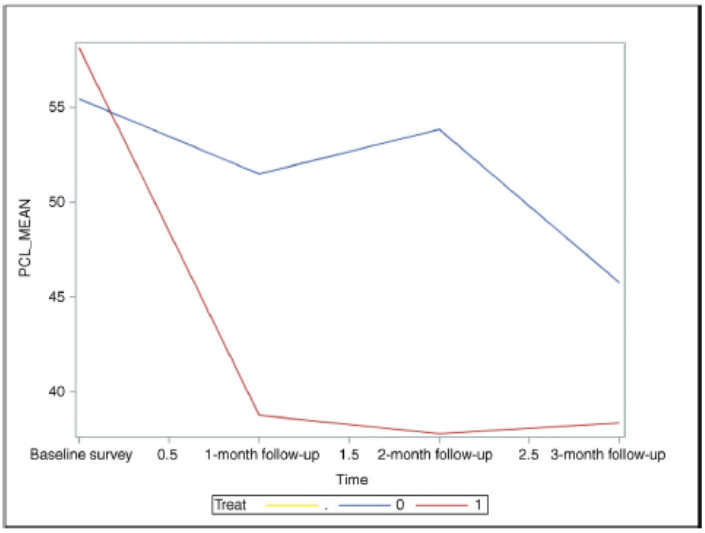


FIGURE 2.2 Time Plot of Mean PCL Scores for Treatment and Control Groups

● Paired T-test

Let $\overline{Y}_{pre}$ and $\overline{Y}_{post}$ be the mean scores of the response before and after a medical treatment for a sample of N patients. If the researcher wants to test whether the two mean scores are different, the null and the alternative hypotheses should be written as $H_0 : \overline{Y}_{post} = \overline{Y}_{pre}$ and $H_1 : \overline{Y}_{post} \neq \overline{Y}_{pre}$, respectively, for a two-tail test. Suppose that the variances of the pre- and posttest mean scores are equal with the same sample size. The equation of the paired t-test is

$$t = \frac{\overline{Y}_{post} - \overline{Y}_{pre}}{s_D / \sqrt{N}}, \qquad (2.1)$$

where s_D is the sample standard deviation of the differences between all pre- and posttest pairs, and the denominator on the right of the equation is the corresponding standard error. The above statistic asymptotically follows a Student's t distribution.

(1) In summary, the paired *t*-test provides a **robust statistical test** to determine the equality or difference of two means measured at two time points. It requires normality of the sample means and a scaled chi-square distribution for sample variances, with the two distributions assumed to be statistically independent. Based on the **central limit theorem**, the sample means usually tend to a normal distribution in probability even if the data are not normally distributed, as long as the sample size is large.

(2) When the longitudinal data deviates markedly from normality and/or the sample size is small, the paired *t*-test should be replaced by the Wilcoxon rank test for paired samples. If longitudinal data includes a large number of follow-up time points, the application of the paired *t*-test is not plausible to test a series of paired mean differences.

● Repeated measures ANOVA

The analysis of change is a fundamental component of so many research endeavors in almost every discipline. Many of the earliest statistical methods for the analysis of change were based on the **analysis of variance (ANOVA)** paradigm, as originally developed by R. A. Fisher. One of the earliest methods proposed for analyzing longitudinal data was a **mixed-effects ANOVA, with a single random subject effect.** The mixed-effects ANOVA model has a long history of use for analyzing longitudinal data, where it is often referred to as the *univariate* repeated- measures ANOVA. Statisticians recognized that a longitudinal data structure, **with *N* individuals and *n* repeated measurements**, has striking similarities to data collected in a **randomized block design (RBD)**, or the closely related **split-plot design**. So it seemed natural to apply ANOVA methods developed for these designs (e.g., Yates, 1935; Scheff´e, 1959) to the repeated-measures data collected from longitudinal studies.

(1) The univariate repeated-measures ANOVA model

- **Its restrictive assumptions:** "compound symmetry" structure for the covariance: constant variance and constant covariance.
- A natural generalization of **Student's (1908) paired *t*-test** to handle more than two repeated measurements, in addition to various between-subject factors.
- **Unbalanced data? =>Missing data problem!**

Q: The repeated-measures *multivariate* analysis of variance (MANOVA) VS Longitudinal Data Analysis ?

As originally developed, MANOVA was intended for the simultaneous analysis of a single measure of a multivariate vector of substantively *distinct* response variables. In contrast, while longitudinal data are multivariate, the vector of responses are commensurate, being repeated measures of the same response variable over time.

- However, MANOVA had a number of features that also limited its usefulness. In particular, the MANOVA formulation forced the **within-subject covariates to be the same for all individuals.**
- Repeated-measures MANOVA cannot be used when the design is **unbalanced over time** (i.e., when the vectors of repeated measures are of different lengths and/or obtained at different sequences of time).
- The repeated-measures MANOVA (at least as implemented in existing statistical software packages) did not **allow for general missing-data patterns** to arise. Thus, if any individual has even a single missing response at any occasion, the entire data vector from that individual must be excluded from the analysis. This so-called "listwise" deletion of missing data from the analysis often results in dramatically reduced sample size and **very inefficient** use of the available data. Listwise deletion of missing data can also produce **biased estimators** of change in the mean response over time when the so-called "completers" (i.e., those with no missing data) are not a random sample from the target population.

In many longitudinal studies there is considerable variation among individuals in both the number and timing of measurements. The resulting data are highly unbalanced and not readily amenable to ANOVA methods developed for balanced designs. It was these features of longitudinal data that provided the impetus for statisticians to develop far more versatile techniques that can handle the commonly encountered problems of data that are **unbalanced and incomplete**, mistimed measurements, time-varying and time-invariant covariates, and responses that are discrete rather than continuous.

When the longitudinal response is discrete, linear models (e.g., linear mixed -effects models) are not very appealing for relating changes in the mean response to covariates for at least two main reasons. First, with a discrete response there is intrinsic **dependence of the variability on the mean.** Second, the **range of the mean response** (a proportion or rate for a response that is binary or a count, respectively) is constrained. In the setting of regression modeling of a univariate response, both of these aspects of the response can be conveniently accommodated within generalized linear models via known variance and link functions. However, a straightforward application of generalized linear models to longitudinal data is not appropriate, **due to the lack of independence among repeated measures obtained on the same individual.**

Advantages and Disadvantages of Longitudinal Data:

Advantages:

1. Although time effects can be investigated in cross-sectional studies in which different subjects are examined at different time points, only longitudinal data give information on individual patterns of change.
2. In investigating time effects in a longitudinal design or treatment effects in a crossover design, each subject can "serve as his or her own control". That is, comparisons can be made within a subject rather than between subjects. This eliminates between-subjects sources of variability from the experimental error and makes inferences more efficient/powerful (think paired *t*-test versus two-sample *t*-test).
3. Since the same variables are measured repeatedly on the same subjects, the reliability of those measurements can be assessed, and purely from a measurement standpoint, reliability is higher.

Disadvantages:

1. For longitudinal or, more generally, **clustered data it is typically reasonable to assume independence across clusters, but repeated measures within a cluster are almost always correlated**, which complicates the analysis.

2. Clustered data are often unbalanced or partially **incomplete (involve missing data)**, which also complicates the analysis. For longitudinal data, this may be due to loss to follow-up (some subjects move away, die, miss appointments, etc.). For other types of clustered data, **the cluster size may vary** (e.g., familial data, where family size varies).