Modeling Data Using Generalized Estimating Equations

October 22, 2017

Outline

GEEs

What are generalized estimating equations?

Models that can be used to estimate relationships between a response and covariate that may not be linear as in generalized linear models where observations are not necessarily independent/

When are they useful?

When the variance/covariance structure is unknown - longitudinal data, clustered data

How do I interpret results?

Similar to linear/generalized linear models

Sleep Study

Test a new insomnia medication:

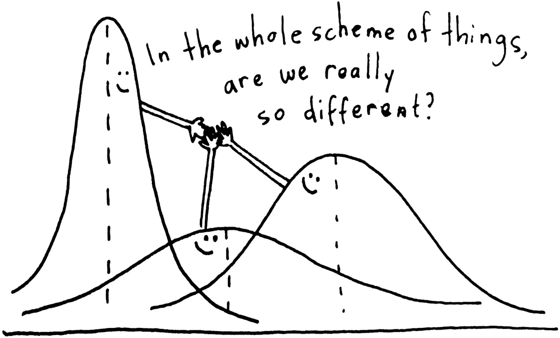
- Experiment: Randomly assign subjects to control and treatment groups
- Explanatory variable = treatment or control group
- What is response?
  - Length of time subject stays asleep
  - Number of hours subject stays asleep
  - Whether or not the subject slept 8 hours (Yes or No)

Typical Linear Model

$$Y = \beta_1 + \beta_2 X + \epsilon$$

- Y is the response variable
- X is the independent variable
- ϵ is the error term usually assumed to be normally distributed

Compare Groups



Generalized Linear Models

If Y is not continuous, normal errors do not make sense:

- Y is a count (number of hours asleep) - Poisson model
- Y is binary (Yes of No) - logistic regression

Clustered Data

Now suppose that the sampling unit is households:

- Randomly assign households to treatment and control.
- Measure response on all household members.
- Responses are correlated within households.
- Confounding effects.
- Error terms are no longer independent.
- Linear models can produce misleading results.

Longitudinal Data

Or suppose that information is obtained on the same individuals at different times during the study:

- Could be a time component - add term to the model
- Measured responses on same individual are correlated
- Linear models can produce misleading results.

Generalized Estimating Equations

- Can use a non-linear link function (count data, binary data).
- Takes correlation within measurements into account.

GEE

$$U(\beta) = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta} \mathbf{V}_i^{-1} (Y_i - \mu_i(\beta))$$

- $\mathbf{Y}_i$ is the response vector of the ith clustered observations - The variables in $\mathbf{Y}_i$ are NOT independent but $\mathbf{Y}_i$ is independent of $\mathbf{Y}_j$
- $\mu(\beta)$ is the mean of Y assumed to be a function of unknown parameters β
- $\mathbf{V}_i$ is the covariance matrix of the ith cluster
- $\mathbf{U}(\beta)$ is the generalized estimating equation

Linear Predictor

The linear predictor is the same as the link function in GLM - it is designed to map the linear predictor into the parameter space.

- $\mu_i(\boldsymbol{\beta}) = \beta_1 + \beta_2 X$ continuous data
- $\log(\mu_i(\boldsymbol{\beta})) = \beta_1 + \beta_2 X$ count data (Poisson)
- $\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_1 + \beta_2 X$ for binary data

Covariance Matrix

Called a working covariance matrix since the exact structure is usually unknown.

- Independent - no correlation
- Exchangeable - Clustered data (order of observations not important)
- Auto correlation - Longitudinal data (time order matters)
- Unstructured - Form unknown

What Covariance Structure to Choose?

Good News!
The parameter estimates are consistent even when the covariance matrix is misspecified!

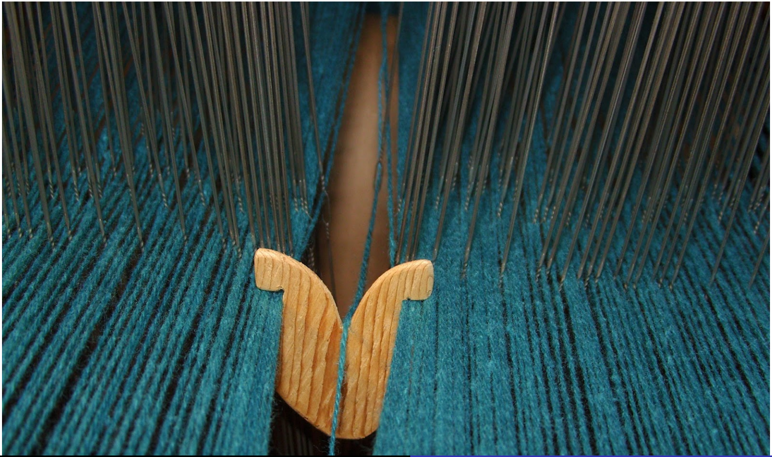
R

gee(formula, id, data,family=gaussian, corstr="independence"....)

- formula is the variables in the linear predictor
- id is the grouping variable
- family is the link function
- corstr is the correlation structure

Warp Break Data

This data set gives the number of warp breaks per loom, where a loom corresponds to a fixed length of yarn. 4.jpg



Warp Break Example

	breaks	wool	tension
1	26	A	L
2	30	A	L
3	54	A	L
4	25	A	L
5	70	A	L
6	52	A	L

Wool type and tension are factors so a two way ANOVA could be performed. But suppose the wool variable is actually a clustering variable so that the breaks with the same type of wool are not independent.

Warp Break Model

Assuming a linear link function:
gee(breaks ~ tension, id=wool, data=warpbreaks,
corstr="exchangeable")

Linear Predictor

(Intercept)	tensionM	tensionH
36.38889	-10.00000	-14.72222

$$\mu = 36.39 - 10 * I_M - 14.72 * I_H$$

- I_M is the indicator for tension level M
- I_H is the indicator for tension level H
- Low tension is considered baseline

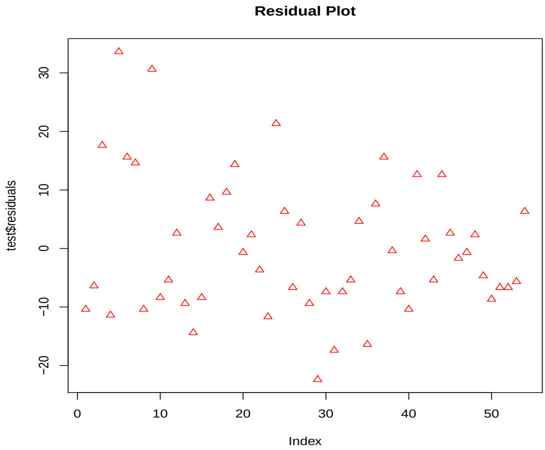
Estimated Mean Levels

Tension Level		
Low	Medium	High
36.39 (5.77)	26.39 (7.46)	21.67 (3.73)

Standard Error of the Means

	(Intercept)	tensionM	tensionH
(Intercept)	33.34722	-43.10185	-21.55093
tensionM	-43.10185	55.70988	27.85494
tensionH	-21.55093	27.85494	13.92747

Diagnostics



Different Working Covariance Matrix

gee(breaks   tension, id=wool, data=warpbreaks,
corstr="AR-M", Mv=1)

Linear Model

	Tension	
Low	Medium	High
36.39 (5.79)	26.39 (7.51)	21.67 (3.76)

Estimates are the same! Standard errors slightly inflated.

Example 2 - Count Data

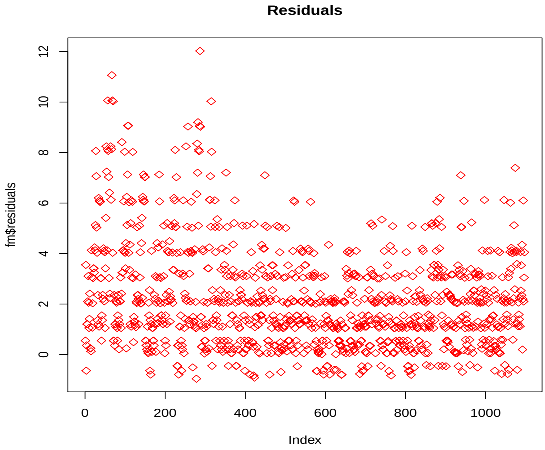
	ID	Age	OME	Loud	Noise	Correct	Trials
1	1	30	low	35	coherent	1	4
2	1	30	low	35	incoherent	4	5
3	1	30	low	40	coherent	0	3
4	1	30	low	40	incoherent	1	1
5	1	30	low	45	coherent	2	4
6	1	30	low	45	incoherent	2	2

Experiments were performed on children on their ability to differentiate a signal in broad-band noise. The noise was played from a pair of speakers and a signal was added to just one channel; the subject had to turn his/her head to the channel with the added signal. The signal was either coherent (the amplitude of the noise was increased for a period) or incoherent (independent noise was added for the same period to form the same increase in power).
The threshold used in the original analysis was the stimulus loudness needs to get 75% correct responses. Some of the children had suffered from otitis media with effusion (OME).

GEE Model

```
gee(cbind(Correct, Trials-Correct) ~ Loud + Age + OME, id = ID,
data = OME, family = binomial, corstr = "exchangeable")
```

	Parameter Estimate	Standard Error	Z-score	p-value
Intercept	-5.901	.231	-25.54	<.001
Loud	0.155	.005	31	<.001
Age	.0185	.003	6.167	<.001
OME High	-0.042	.152	-.276	.391
OME Low	-.286	.118	2.431	.008


Residuals
fm$residuals
0
2
4
6
8
10
12
0
200
400
600
800
1000
Index

In Conclusion

- GEE’s are useful tools for data that are not continuous.
- GEE’s are useful for longitudinal and clustered data
- GEE’s can be used to test for significance
- Diagnostic on GEE model fit