Longitudinal Data: Repeated Measures

Methods for Longitudinal Repeated Measures Data

Alan Hubbard

Oct 16, 2018

Load libraries and read in data

```
dat=read_csv("teensex.csv")
## Look at data
tbl_df(dat)
```

```
## # A tibble: 1,909 x 4
##       eid today     sx24hrs drgalcoh
##     <int> <chr>       <int>    <int>
##  1      1 3-Jun-98        0        1
##  2      2 4-Jun-98        0        0
##  3      2 5-Jun-98        0        0
##  4      2 6-Jun-98        0        1
##  5      2 7-Jun-98        0        0
##  6      2 8-Jun-98        0        0
##  7      2 9-Jun-98        0        0
##  8      2 12-Jun-98       0        0
##  9      2 14-Jun-98       0        1
## 10      2 16-Jun-98       0        0
## # ... with 1,899 more rows
```

Data,Notation

- child = “eid”"
- dat of survey is “today”
- drug or alchol use is drgalcoh, $X_{ij} = 1$ yes, 0 for no.
- outcome (sexual activity) $Y_{ij} = 1$ for yes, 0 for no.

Convert "today" variable into a system-recognized date.

```
knitr::asis_output("\\footnotesize")
```

```
# dmy converts strings in form of day-month-yr into system
# dates. year gets the year from the date.
dat <- dat %>% mutate(date=dmy(today)) %>% mutate(year=year(date))
# get rid of observations (probably errors) with year < 1998
dat <- subset(dat,year > 1997)
# Group by id to get minimum date by id
dat <- dat %>% group_by(eid) %>% mutate(mind = min(date,na.rm=F))
```

Data Summaries

- We do simple summaries to look at the proportion of observations that report sexual activity overall and by id, as well as by drug/alchol group

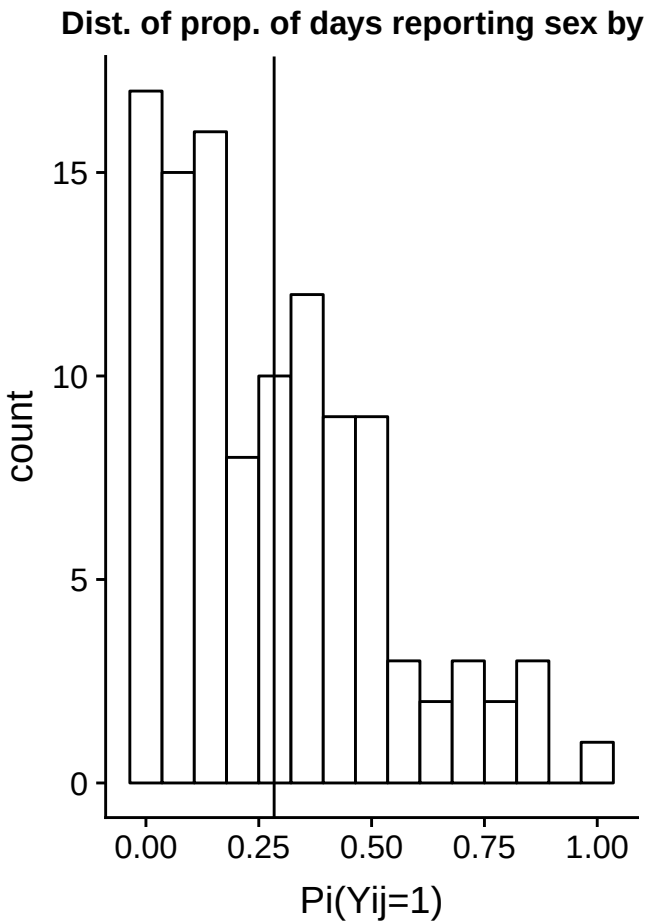
```
## # A tibble: 3 x 2
##   drgalcoh   PY1
##      <int> <dbl>
## 1        0 0.256
## 2        1 0.374
## 3       NA 0.055
```

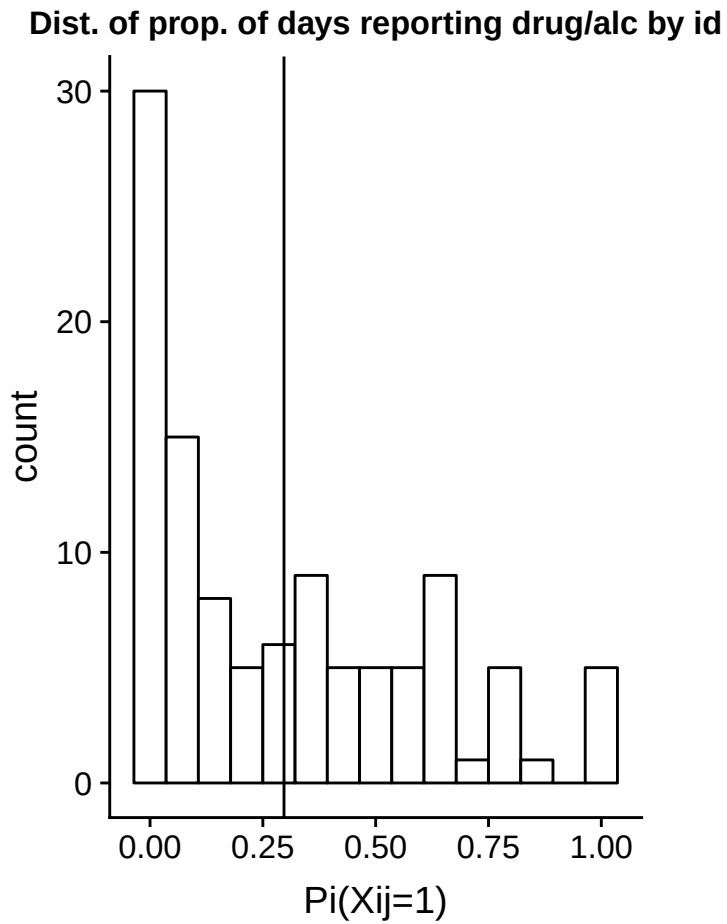
Display plots

```
knitr::asis_output("\\tiny")
```

```
plot_grid(p1,p2, scale=0.7)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```





Comments on data summaries

- One can see large variation in the proportion of days reported for both relevant events (both X_{ij}, Y_{ij}) across individuals.
- Subjects report an average of around 25% yeses for both variables.
- We will look at ways of modeling this variability below.

Transition Models

-We wish to fit the model:

$$logit[P(Y_{ij} = 1 \mid X_{ij}, Y_{i,j-1})] = \beta_0^{TM} + \beta_1^{TM} X_{ij} + \beta_2^{TM} Y_{i,j-1}$$

- We will assume that

$$Y_{ij} \perp (Y_{i1}, Y_{i2}, \cdots, Y_{i,j-2}) \mid X_{ij}, Y_{i,j-1}$$

so, no correlation with past Y's given the most recently measured one.

- First, we need to make a variable of the lagged value of outcome (by 1 row), ** by id **.
- Note first, that because one of the variables depends on the value the next observations, it's blank for the first observation.

Make lag variable

```
knitr::asis_output("\\tiny")
```

```
# order by date with eid
dat <- arrange(dat,eid,date)
# get new variable which is lagged (by 1) outcome (Y_{ii,j-1})
dat_new <- dat %>% group_by(eid) %>% mutate(yprev = lag(sx24hrs))
dat_new[22:33,]
```

```
## # A tibble: 12 x 8
## # Groups:   eid [2]
##       eid today     sx24hrs drgalcoh date        year mind       yprev
##     <int> <chr>       <int>    <int> <date>     <dbl> <date>     <int>
##  1     2 3-Jul-98        0        0 1998-07-03  1998 1998-06-04     0
##  2     2 4-Jul-98        0        0 1998-07-04  1998 1998-06-04     0
##  3     2 5-Jul-98        0        0 1998-07-05  1998 1998-06-04     0
##  4     3 4-Jun-98        0        0 1998-06-04  1998 1998-06-04    NA
##  5     3 7-Jun-98        0        0 1998-06-07  1998 1998-06-04     0
##  6     3 8-Jun-98        0        0 1998-06-08  1998 1998-06-04     0
##  7     3 9-Jun-98        0        0 1998-06-09  1998 1998-06-04     0
##  8     3 10-Jun-98       0        0 1998-06-10  1998 1998-06-04     0
##  9     3 11-Jun-98       0        0 1998-06-11  1998 1998-06-04     0
## 10     3 12-Jun-98       0        0 1998-06-12  1998 1998-06-04     0
## 11     3 13-Jun-98       1        1 1998-06-13  1998 1998-06-04     0
## 12     3 14-Jun-98       0        0 1998-06-14  1998 1998-06-04     1
```

Fit standard logistic regression

- Given our assumptions of conditional independence above, we have no residual correlation of observations on the same subject conditional on the past outcome and current drug and alcohol use.
- Thus, we fit the coefficients using standard logistic regression.

GLM in R

```
knitr::asis_output("\\tiny")
```

```
glm_tm <- glm(sx24hrs ~ drgalcoh+yprev, data=dat_new, family =binomial(),na.action=na.omit)
summary(glm_tm)
```

```
##
## Call:
## glm(formula = sx24hrs ~ drgalcoh + yprev, family = binomial(),
##     data = dat_new, na.action = na.omit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1673  -0.8823  -0.7142   1.1875   1.7269
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.23601    0.07594 -16.276  < 2e-16 ***
## drgalcoh     0.49346    0.12129   4.069 4.73e-05 ***
## yprev        0.71877    0.12079   5.950 2.67e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1940.6  on 1606  degrees of freedom
## Residual deviance: 1885.2  on 1604  degrees of freedom
##   (301 observations deleted due to missingness)
## AIC: 1891.2
##
## Number of Fisher Scoring iterations: 4
```

Get associations at right scale

- See a strong positive relationship of both previous recorded days sexual activity as well as current days reporting alcohol/drug use and sexual activity.
- We now put into more interpretable odds ratio form. Again, can do this several ways, but we'll use the function we've used before. We only have binary predictors, so the relevant OR's are just comparing yes (1) to no (0) for both.

```
knitr::asis_output("\\footnotesize")
```

```
source("glm_post_estimate.R")
comps <- rbind(c(0,1,0),c(0,0,1))
or_tm <- glm.post.estimate(glm_tm,comps,
  labs=c("drug/alc","previousY"),exponentiate = T)
or_tm
```

```
##           Ratio.est CI                pvalue
## drug/alc  "1.638"   "1.291 - 2.078" "0.000"
## previousY "2.052"   "1.619 - 2.600" "0.000"
```

Interpretation

- One can see that the estimate odds ratio suggests about a doubling of the probability of reporting sexual activity if it was reported the previous day (keep drugs and alcohol fixed).
  - This obviously suggest a strong correlation of outcomes at least over a 1-day interval.
- There is also a strong positive association of drug/alc use in the current day versus reported sexual activity.

Random (mixed) effects models

- We will talk extensively about such models in next part of class, but for now a brief introduction using the data on teenagers.
- This approach implicitly models the correlation by assuming a latent variable model that implies correlation of observations made on same subject.
- Specifically in this case, the model is:

$$logit[P(Y_{ij} = 1 \mid X_{ij}, \beta_{0i}] = \beta_0^{RE} + \beta_{0i} + \beta_1^{RE} X_{ij}$$

where it is assumed that the β_{0i} are random, independent normally distributed variables: $\beta_{0i} \sim N(0, \tau)$. - This models assumes that

$$Y_{ij} \perp Y_{ij'}, j' \neq j, \mid \beta_{0i}, X_{ij},$$

.

How to estimate it in R

```
knitr::asis_output("\\tiny")
```

```
glmer_mod <- glmer(sx24hrs ~ drgalcoh+(1|eid), data=dat_new, family =binomial(),na.action=na.omit)
summary(glmer_mod)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: sx24hrs ~ drgalcoh + (1 | eid)
##    Data: dat_new
##
##      AIC      BIC   logLik deviance df.resid
##   1852.1   1868.4   -923.1   1846.1     1705
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8638 -0.5799 -0.3905  0.5950  3.6907
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  eid    (Intercept) 1.554    1.247
## Number of obs: 1708, groups:  eid, 109
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0815     0.1504  -7.189 6.53e-13 ***
## drgalcoh      0.3908     0.1530   2.554   0.0107 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##          (Intr)
```

get other estimates of unique to the mixed models (such as $\tau \equiv SD(\beta_{0i})$)

- Other information besides "fixed effect" coefficients.
- For instance, the standard deviation of the random effects (we just have one in this case).
- We will spend more time on this later.
- Below is the estimate of $SD(\beta_{0i})$

```
VarCorr(glmer_mod)
```

```
##  Groups Name        Std.Dev.
##  eid    (Intercept) 1.2465
```

Generalized Estimating Equation (GEE)

- Finally, we fit a simple linear model, but use the GEE approach to get the inference accounting for the repeated measures.
- As discussed in slides, we can use different working correlation matrices to fit the coeffients.
- In this case, we use independence, so the estimates will be the same as if done by standard logistic regression.
- However, the robust inference will adjust for the correlation when deriving standard errors (SE'S).

GEE in R

```
knitr::asis_output("\\tiny")
```

```
gee_fit <- gee(sx24hrs ~ drgalcoh, eid, data=dat_new,family=binomial, corstr = "independence")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate

## (Intercept)    drgalcoh
##   -1.067938    0.553610
```

```
# Make easier to read summary
ss <- data.frame(summary(gee_fit)$coefficients)
ss = data.frame(ss,pvalue=2*(1-pnorm(abs(ss[,5]))))
round(ss,4)
```

```
##             Estimate Naive.S.E.  Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)  -1.0679     0.0648 -16.4707      0.1100  -9.7085 0.0000
## drgalcoh      0.5536     0.1164   4.7543      0.1802   3.0714 0.0021
```

We need slightly altered functions to get post-hoc estimates of linear combination of coefficients.

```
knitr::asis_output("\\tiny")
```

```
source("gee_post_estimate.R")
comps <- c(0,1)
or_gee <- gee.post.estimate(gee_fit,comps,labs=c("drug/alc"),exponentiate = T)
or_gee
```

Comparing GEE/RE model

- One can see that, though both the GEE approach and the random effects model account for correlation when deriving their inference, they give quite different estimates of the the OR (1.478 versus 1.740).
- Why??

Different correlation model for GEE

- Now we do a GEE model again, but using a different working correlation model (exchangeable in this case).
- This one assumes that all observations on the same individual are equally correlated.
- It just involves one small change in the gee function.

Re-fit with exchangeable working correlation

```
knitr::asis_output("\\tiny")
```

```
gee_fit <- gee(sx24hrs ~ drgalcoh, eid, data=dat_new,family=binomial, corstr = "exchangeable")
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27
```

```
## running glm to get initial regression estimate
```

```
## (Intercept)    drgalcoh
##   -1.067938    0.553610
```

```
# Make easier to read summary
ss <- data.frame(summary(gee_fit)$coefficients)
ss = data.frame(ss,pvalue=2*(1-pnorm(abs(ss[,5]))))
round(ss,4)
```

```
##             Estimate Naive.S.E. Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)  -0.8751     0.1071 -8.1700      0.1163  -7.5274 0.0000
## drgalcoh      0.3321     0.1188  2.7941      0.1371   2.4219 0.0154
```

```
or_gee_exch <- gee.post.estimate(gee_fit,comps,labs=c("drug/alc"),exponentiate = T)
or_gee_exch
```

```
##          Ratio.est CI              pvalue
## drug/alc "1.394"   "1.065 - 1.824" "0.015"
```

Comparing independence and exchangeable working correlation

- As one can see, there is also a big difference in the estimate of the odds ratio of drug/alcohol for two different runs of the GEE approach, but with different working correlation models.
- Why?