Longitudinal Data: Repeated Measures

Linear Models with repeated measures

Alan Hubbard

Oct 2, 2018

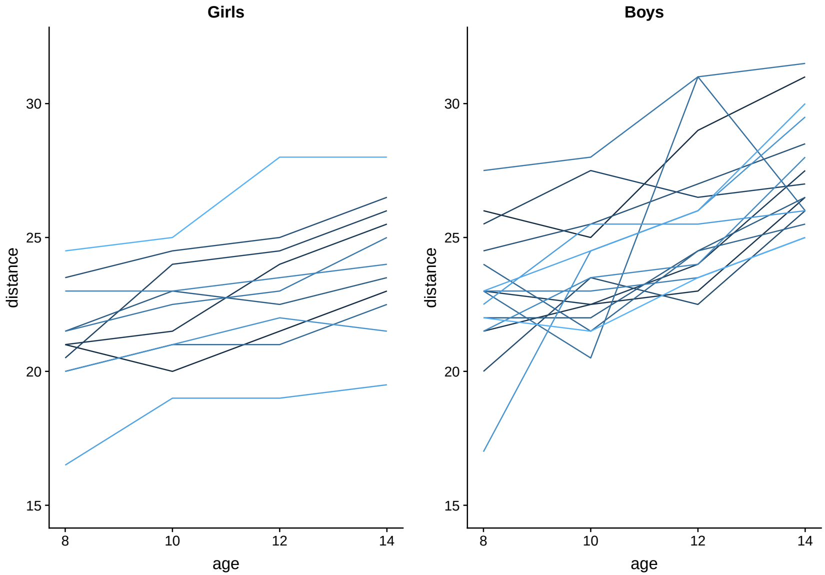
```r
dat <- read_csv("dental.csv")
tbl_df(dat)
```

```
## # A tibble: 108 x 5
##    obsno child   age distance gender
##    <int> <int> <int>    <dbl>  <int>
##  1     1     1     8     21        0
##  2     2     1    10     20        0
##  3     3     1    12     21.5      0
##  4     4     1    14     23        0
##  5     5     2     8     21        0
##  6     6     2    10     21.5      0
##  7     7     2    12     24        0
##  8     8     2    14     25.5      0
##  9     9     3     8     20.5      0
## 10    10     3    10     24        0
## # ... with 98 more rows
```

Data Notation

- child = i (id)
- *age* $= X_{ij1}$
- *gender* $= X_{ij2}$: 1 is male, 0 female
- measure of teeth *distance* $= Y_{ij}$

Plot the data



Look at basic summaries with dplyr

```
# get number of obs per person and mean
sum_dat=group_by(dat,child) %>% summarise(mean=mean(distanc
table(sum_dat$gender)
```

```
##
##  0  1
## 11 16
```

Estimate linear model for dental data

- Again, the model we wanted to estimate is:

$$E(Y_{ij} \mid \vec{X}_{ij}) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij1} X_{ij2}$$

Use simple OLS

```r
# First, make interaction term
library(RCurl)
dat = dat %>% mutate(inter=gender*age)
ols_fit = lm(distance ~ gender+age+inter, data=dat)
summary(ols_fit)
```

```
##
## Call:
## lm(formula = distance ~ gender + age + inter, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6156 -1.3219 -0.1682  1.3299  5.2469
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.3727     1.7080  10.171  < 2e-16 ***
## gender       -1.0321     2.2188  -0.465  0.64279
## age           0.4795     0.1522   3.152  0.00212 **
## inter         0.3048     0.1977   1.542  0.12608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.257 on 104 degrees of freedom
## Multiple R-squared:  0.4227, Adjusted R-squared:  0.4061
## F-statistic: 25.39 on 3 and 104 DF,  p-value: 2.108e-12
```

Getting estimates of linear combination of coefficients

```
## Note how the algebra is entered into function
### For boys
lin.boys=glht(ols_fit, linfct = c("6*age +6*inter=0"))
#summary(lin.boys)
confint(lin.boys)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = distance ~ gender + age + inter, data = dat)
##
## Quantile = 1.983
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                         Estimate lwr    upr
## 6 * age + 6 * inter == 0 4.7063   3.2051 6.2074
```

Repeat for girls

In this case,

$$E(Y_{ij} \mid X_{ij1} = age + 6, X_{ij2} = 0) - E(Y_{ij} \mid X_{ij1} = age, X_{ij2} = 0)$$

$$= 6 * \beta_1$$

```
lin.girls=glht(ols_fit, linfct = c("6*age=0"))
#summary(lin.girls)
confint(lin.girls)
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = distance ~ gender + age + inter, data = dat)
##
## Quantile = 1.983
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##              Estimate lwr    upr
## 6 * age == 0 2.8773   1.0668 4.6877
```

The results suggest that the mean change in the distance for boys is 4.76 (95% 3.20-6.21) and that for girls is 2.88 (95% 1.07-4.69).

Inference sandwich estimator

```r
# Read in user written function using lmtest and
# sandwich packages that return both information
# on coefficients and estimates of linear
# combination of coefficients using
# clustered/robust variance estimates
source("clx.R")
source("clx_lincom.R")

clx(ols_fit, 1, dat$child)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 17.372727   0.749604 23.1759 < 2.2e-16 ***
## gender      -1.032102   1.424137 -0.7247   0.47025
## age          0.479545   0.065257  7.3486 4.712e-11 ***
## inter        0.304830   0.120799  2.5234   0.01313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Robust estimate for change in mean
# for 6 year increase in age for
# boys
clx.lincom(ols_fit, 1,dat$child, c(0,0,6,6))
```

```
##      Est      CI                  pvalue
## [1,] "4.7063" "3.5108 - 5.9017" "0.0000"
```

Comparison of results

- If one looks at the age coefficient, one can see that the robust SE is much smaller than the naive one returned by standard OLS: 0.065 vs 0.152.
- We can also compare the results of the estimate of

$$= 6 * \beta_1 + 6 * \beta_3$$

and we see the CI based upon the robust estimates that account for dependence is 3.511-5.911 versus 3.205-6.207, so for this parameter, the CI gets bigger when the SE is estimated correctly.