Longitudinal Data: Repeated Measures

And More GEE

Alan Hubbard

Oct 23, 2018

Intro

- The Six Cities Study of Air Pollution and Health was a longitudinal study designed to characterize lung growth as measured by changes in pulmonary function in children and adolescents.
- A cohort of 13,379 children born on or after 1967 wasn enrolled in six communities across the U.S.: Watertown (Massachusetts), Kingston and Harriman (Tennessee), a section of St. Louis (Missouri), Steubenville (Ohio), Portage (Wisconsin), and Topeka (Kansas).

Intro, cont.

- Most children were enrolled in the first or second grade (between the ages of six and seven) and measurements of study participants were obtained annually until graduation from high school or loss to follow-up.
- At each annual examination, spirometry, the measurement of pulmonary function, was performed and a respiratory health questionnaire was completed by a parent or guardian.
- Reference: Dockery, D.W., Berkey, C.S., Ware, J.H., Speizer, F.E. and Ferris, B.G. (1983). Distribution of FVC and FEV1 in children 6 to 11 years old. American Review of Respiratory Disease, 128, 405-412.

Specific data for example

- The data consist of all measurements of FEV1, height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas.
- The random sample consists of 299 girls, with a minimum of one and a maximum of twelve observations over time.
- Variables include time-dependent age (yrs), height (meters), and fev1 (forced expired volume in first second after spirometry in ml).
- Data already has been processed somewhat to have baseline age and height as covariates.

Read in data

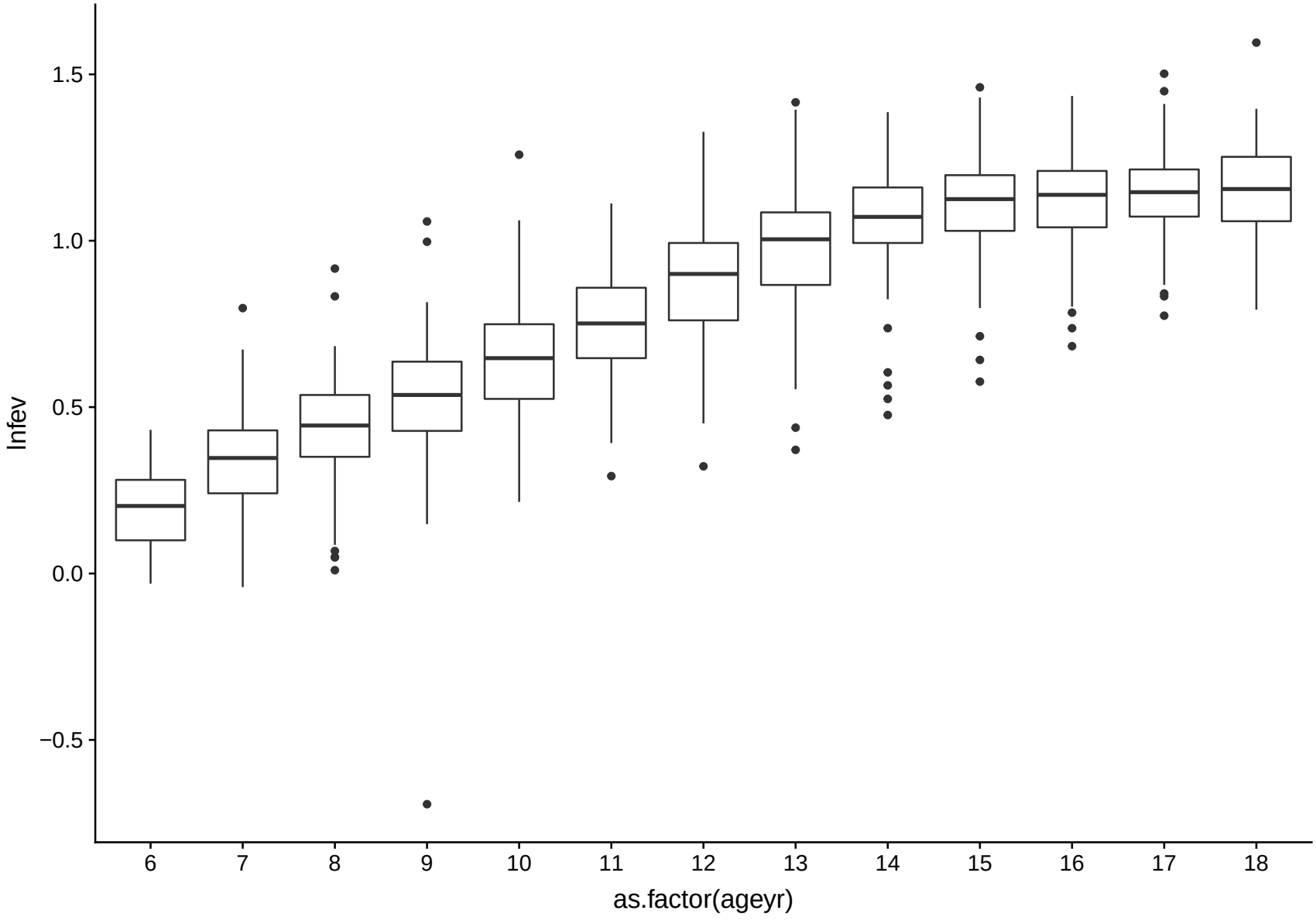
```
dat1 = read_csv("fev.csv")
```

```
## Parsed with column specification:
## cols(
##   childid = col_integer(),
##   id = col_integer(),
##   height = col_double(),
##   age = col_double(),
##   initht = col_double(),
##   initage = col_double(),
##   lnfev = col_double(),
##   lnheight = col_double(),
##   initlnheight = col_double(),
##   agechange = col_double(),
##   lnheightchange = col_double(),
##   year = col_integer()
## )
```

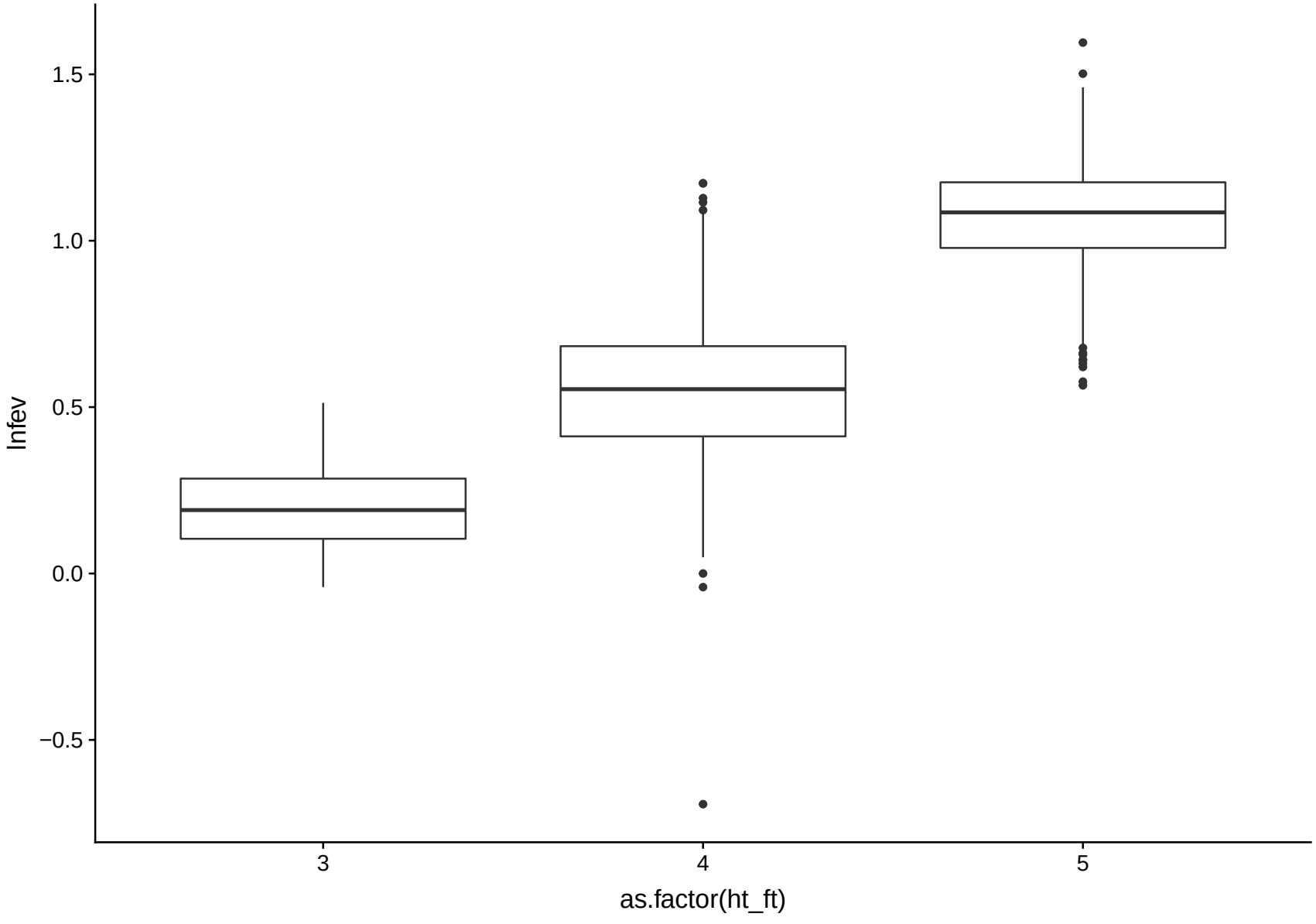
```
tbl_df(dat1)
```

```
## # A tibble: 1,994 x 12
##    childid    id height   age initht initage lnfev lnheight initlnheight
##      <int> <int>  <dbl> <dbl>  <dbl>   <dbl> <dbl>    <dbl>        <dbl>
##  1       1    59   1.2   9.34   1.2     9.34 0.215    0.182        0.182
##  2       1    59   1.28 10.4    1.2     9.34 0.372    0.247        0.182
##  3       1    59   1.33 11.5    1.2     9.34 0.489    0.285        0.182
##  4       1    59   1.42 12.5    1.2     9.34 0.751    0.351        0.182
##  5       1    59   1.48 13.4    1.2     9.34 0.833    0.392        0.182
##  6       1    59   1.5  15.5    1.2     9.34 0.892    0.405        0.182
##  7       1    59   1.52 16.4    1.2     9.34 0.871    0.419        0.182
##  8       2   159   1.13  6.59   1.13    6.59 0.307    0.122        0.122
##  9       2   159   1.19  7.65   1.13    6.59 0.351    0.174        0.122
## 10       2   159   1.49 12.7    1.13    6.59 0.756    0.399        0.122
```

Distribution of FEV1 by Age



Distribution of FEV1 by Height in Feet



Review of longitudinal versus cross-sectional effects

Consider the model:

$$E(Y_{ij} \mid X_{ij}, X_{i1}) = \beta_0 + \beta_1 X_{i1} + \beta_2 (X_{ij} - X_{i1})$$

- In this case, β_1 is the so-called cross-sectional association, where as β_2 is the longitudinal (a change in the mean for a one unit change in covariate from baseline, keeping baseline value fixed).

Alternative parameterization

- An equivalent (will fit the data exactly the same) is

$$E(Y_{ij} \mid X_{ij}, X_{i1}) = \beta_0 + \beta_1^* X_{i1} + \beta_2 X_{ij}$$

  where we know use β_1^* instead of β_1 because they are different parameters.
- However, it is easy to see that $\beta_1^* = \beta_1 - \beta_2$.
- Thus, it's just a question of which output is more conventient since these are both equivalent (much like choosing different baseline values in making dummy variables does not change the fit nor estimation of same parameters).

Our specific models

- We will use just one of these parameterizations, but not just with one variable, but two (height and age).
- Thus, our models become either (with X_{ij1} log(height) for subject i, time j, and X_{ij2} is the corresponding age):

$$E(Y_{ij} \mid X_{ij1}, X_{i11}, \mid X_{ij2}, X_{i12}) = \beta_0 + \beta_1 X_{i11} + \beta_2 (X_{ij1} - X_{i11}) +$$

$$\beta_3 X_{i12} + \beta_4 (X_{ij2} - X_{i12})$$

Independence Working Correlation

```r
# independence

gee_ind <- gee(lnfev ~ +initlnheight + lnheightchange + initage +
    agechange, childid, data = dat1, family = gaussian, corstr = "independence")
```

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

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

```
##    (Intercept)   initlnheight lnheightchange        initage      agechange
##    -0.33093749     2.46367971     2.05618307     0.01241088     0.02849790
```

```r
# Make easier to read summary
ss1 <- data.frame(summary(gee_ind)$coefficients)
ss1 = data.frame(ss1, pvalue = 2 * (1 - pnorm(abs(ss1[, 5]))))
round(ss1, 4)
```

```
##                Estimate Naive.S.E.  Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)     -0.3309     0.0211 -15.7018      0.0432  -7.6616 0.0000
## initlnheight     2.4637     0.0651  37.8573      0.1772  13.9000 0.0000
## lnheightchange   2.0562     0.0700  29.3738      0.0792  25.9775 0.0000
## initage          0.0124     0.0034   3.6075      0.0087   1.4202 0.1555
## agechange        0.0285     0.0021  13.4835      0.0023  12.5447 0.0000
```

Results

Comparison of Standard Errors

Variable	Naïve SE	Robust SE	Naïve z	Robust z
lnheight	.0699	.0793	29.4	25.9
age	.0021	.0023	13.5	12.5
initlnheight	.0840	.1829	4.8	2.2
initage	.0040	.0088	-4.0	-1.8
_cons	.0211	.0433	-15.7	-7.6

Get estimates of changes in age and changes in height

```
source("gee_post_estimate.R")
comps = rbind(c(0, 0, 0.5, 0, 0), c(0, 0, 0, 0, 1))
gee.post.estimate(gee_ind, comps, labs = c("log(height)", "
    rounded = 3, exponentiate = FALSE)
```

```
##                Est     CI                  pvalue
## log(height) "1.028" "0.951 - 1.106" "0.000"
## age         "0.028" "0.024 - 0.033" "0.000"
```

Exchangeable

```
gee_exc <- gee(lnfev ~ +initlnheight + lnheightchange + initage +
    agechange, childid, data = dat1, family = gaussian, corstr = "exchangeable")
```

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

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

```
##    (Intercept)   initlnheight lnheightchange        initage      agechange
##    -0.33093749     2.46367971     2.05618307     0.01241088     0.02849790
```

```
# Make easier to read summary
ss2 <- data.frame(summary(gee_exc)$coefficients)
ss2 = data.frame(ss2, pvalue = 2 * (1 - pnorm(abs(ss2[, 5]))))
round(ss2, 4)
```

```
##                Estimate Naive.S.E. Naive.z Robust.S.E. Robust.z pvalue
## (Intercept)     -0.2812     0.0389 -7.2299      0.0450  -6.2562 0.0000
## initlnheight     2.6278     0.1406 18.6954      0.1937  13.5683 0.0000
## lnheightchange   2.1979     0.0466 47.1447      0.0467  47.0517 0.0000
## initage          0.0010     0.0073  0.1295      0.0099   0.0959 0.9236
## agechange        0.0243     0.0014 17.4488      0.0013  18.9640 0.0000
```

Get estimates of changes in age and changes in height

```
comps = rbind(c(0, 0, 0.5, 0, 0), c(0, 0, 0, 0, 1))
gee.post.estimate(gee_exc, comps, labs = c("log(height)", "
    rounded = 3, exponentiate = FALSE)
```

```
##                Est     CI                  pvalue
## log(height) "1.099" "1.053 - 1.145" "0.000"
## age         "0.024" "0.022 - 0.027" "0.000"
```