Week 6 Practical Session

David Barron

Trinity Term 2018

Longitudinal data analysis

Many countries have a longitudinal household survey. In the UK, the British Household Panel Survey (now continued by a new survey called Understanding Society) is such a survey. They make available a teaching version of the data. It's simplified, but still much more complex than anything we've used before. There are 9,912 cases (to keep it simpler, this version of the data only includes people included in wave 1) and 760 variables. The data are in wide format. BHPS uses a letter prefix to identify the wave the variable was collected, so for example the variable that indicates whether someone works full time or part time is called `ajbft` in wave 1, `bjbft` in wave 2, and so on.

The difficult part is turning this into the long format of data that is needed for analysis. The only way to do it is to select one set of variables at a time, turn that set of variables into a single long format variable along with an id variable and a wave variable. When you have done that for all the variables you want to include in the long data set, you then have to merge them together.

First, read the data in and select the variables that don't change over time.

```
bhps <- foreign::read.dta("C:\\Users\\dbarron\\Dropbox\\Advanced Quant\\BHPS\\stata9\\bhps_sampler3.dta
dim(bhps)
```

```
[1] 9912  760
```

```
bhps1 <- dplyr::select(bhps, pid, sex, aage)
```

To select variables it makes sense to create a small function to do the work. Then this can be used to select as many variables as we want.

```
selvar <- function(vn) {
    require(tidyr, quietly = TRUE, warn.conflicts = FALSE)
    require(dplyr, quietly = TRUE, warn.conflicts = FALSE)
    require(stringr, quietly = TRUE, warn.conflicts = FALSE)

    op <- bhps %>% select(pid, ends_with(vn)) %>% gather(wave, variable, -pid) %>%
        mutate(wave = str_sub(wave, 1, 1))

    names(op)[3] <- vn

    op
}

bhps2 <- selvar("jbstat")
bhps3 <- selvar("jbsect")
bhps4 <- selvar("fiyr")
bhps5 <- selvar("jbft")
bhps6 <- selvar("mlstat")
bhps7 <- selvar("vote")
```

Then we can merge these to create a new data set. I'm using a function called `inner_join` from the `dplyr` package. There is a function called `merge` in base R that you could use instead if you prefer.

```r
bhps_sub <- inner_join(bhps1, bhps2, by = c("pid"))
bhps_sub <- inner_join(bhps_sub, bhps3, by = c("pid", "wave"))
bhps_sub <- inner_join(bhps_sub, bhps4, by = c("pid", "wave"))
bhps_sub <- inner_join(bhps_sub, bhps5, by = c("pid", "wave"))
bhps_sub <- inner_join(bhps_sub, bhps6, by = c("pid", "wave"))
bhps_sub <- inner_join(bhps_sub, bhps7, by = c("pid", "wave"))
names(bhps_sub)
```

```
 [1] "pid"    "sex"    "aage"   "wave"   "jbstat" "jbsect" "fiyr"  
 [8] "jbft"   "mlstat" "vote"  
```

```r
dim(bhps_sub)
```

```
[1] 118944     10
```

```r
xtabs(~wave, bhps_sub)
```

```
wave
   a    c    d    e    f    g    h    i    j    k    l    m 
9912 9912 9912 9912 9912 9912 9912 9912 9912 9912 9912 9912 
```

Now we need to clean up the data. Again, I've created a short function to create missing data codes.

```r
toNA <- function(var, lv) {
    # Turn selected values into NA
    var[var %in% lv] <- NA
    # Output transformed variables, dropping an unused factor levels
    var[, drop = TRUE]
}

# Turn character variables into factors
bhps_sub$jbsect <- factor(bhps_sub$jbsect)
bhps_sub$jbstat <- factor(bhps_sub$jbstat)
bhps_sub$wave <- factor(bhps_sub$wave)
bhps_sub$jbft <- factor(bhps_sub$jbft)

# Missing data

bhps_sub$employed <- toNA(bhps_sub$jbstat, levels(bhps_sub$jbstat)[c(1, 12,
    15, 19)])

bhps_sub$ft <- toNA(bhps_sub$jbft, levels(bhps_sub$jbft)[c(2, 3, 5)])
# Create a numeric wave variable
bhps_sub$wavenum <- match(bhps_sub$wave, letters)
# Use this to create age
bhps_sub$age <- bhps_sub$aage + bhps_sub$wavenum - 1
# Create log income variable
bhps_sub$fiyr[bhps_sub$fiyr <= 0] <- NA

bhps_sub$logfiyr <- log(bhps_sub$fiyr)


bhps_sub$ft <- factor(bhps_sub$ft)

bhps_sub$vote <- forcats::fct_recode(factor(bhps_sub$vote), `NULL` = "Can't vote",
    `NULL` = "Don't know", `NULL` = "Missing or wild", `NULL` = "Proxy and or phone",
    `NULL` = "Proxy respondent", `NULL` = "Refused", `NULL` = "Respondent absent this wave",
    Other = "Other Party", Other = "Other answer")

bhps_sub <- bhps_sub %>% arrange(pid, wavenum) %>% group_by(pid) %>% mutate(start_ft = ifelse(ft ==
    "Full time: 30 hrs +" & lag(ft) == "Part time: lt 30 hrs", "Yes", "No"),
    ch_vote = ifelse(vote == lag(vote), FALSE, TRUE), ch_inc = (fiyr - lag(fiyr))/lag(fiyr))

# ggplot(bhps_sub, aes(x = fiyr)) + geom_density()
```

Now we are ready to do some analysis! First, simple random effects and fixed effects models.

```r
p1 <- plm(logfiyr ~ sex + ft + age + I(age^2/1000), data = bhps_sub, index = c("pid",
    "wavenum"))

summary(p1)
```

```
Oneway (individual) effect Within Model

Call:
plm(formula = logfiyr ~ sex + ft + age + I(age^2/1000), data = bhps_sub,
    index = c("pid", "wavenum"))

Unbalanced Panel: n = 6882, T = 1-12, N = 48647

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-9.666481 -0.118127  0.013057  0.174875  6.878609

Coefficients:
                         Estimate Std. Error t-value  Pr(>|t|)
ftPart time: lt 30 hrs -0.4232457  0.0106967 -39.568 < 2.2e-16 ***
age                     0.1636588  0.0031138  52.559 < 2.2e-16 ***
I(age^2/1000)          -1.2997784  0.0370391 -35.092 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    15883
Residual Sum of Squares: 13242
R-Squared:      0.16631
Adj. R-Squared: 0.02889
F-statistic: 2777.06 on 3 and 41762 DF, p-value: < 2.22e-16
```

```r
p2 <- plm(logfiyr ~ sex + ft + age + I(age^2/1000), data = bhps_sub, index = c("pid",
    "wavenum"), model = "random")
summary(p2)
```

```
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = logfiyr ~ sex + ft + age + I(age^2/1000), data = bhps_sub,
    model = "random", index = c("pid", "wavenum"))

Unbalanced Panel: n = 6882, T = 1-12, N = 48647
```

```
Effects:
                 var std.dev share
idiosyncratic 0.3171  0.5631 0.548
individual    0.2614  0.5113 0.452
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2596  0.6371  0.6848  0.6469  0.6970  0.6970

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-9.7253 -0.1626  0.0692  0.0143  0.2817  3.4240

Coefficients:
                         Estimate Std. Error t-value  Pr(>|t|)
(Intercept)             6.4343826  0.0515153 124.902 < 2.2e-16 ***
sexFemale              -0.4110254  0.0149640 -27.468 < 2.2e-16 ***
ftPart time: lt 30 hrs -0.5318588  0.0099616 -53.391 < 2.2e-16 ***
age                     0.1348840  0.0024789  54.412 < 2.2e-16 ***
I(age^2/1000)          -1.3063977  0.0289854 -45.071 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    46350
Residual Sum of Squares: 16505
R-Squared:      0.64736
Adj. R-Squared: 0.64733
F-statistic: 21989.8 on 4 and 48642 DF, p-value: < 2.22e-16
```

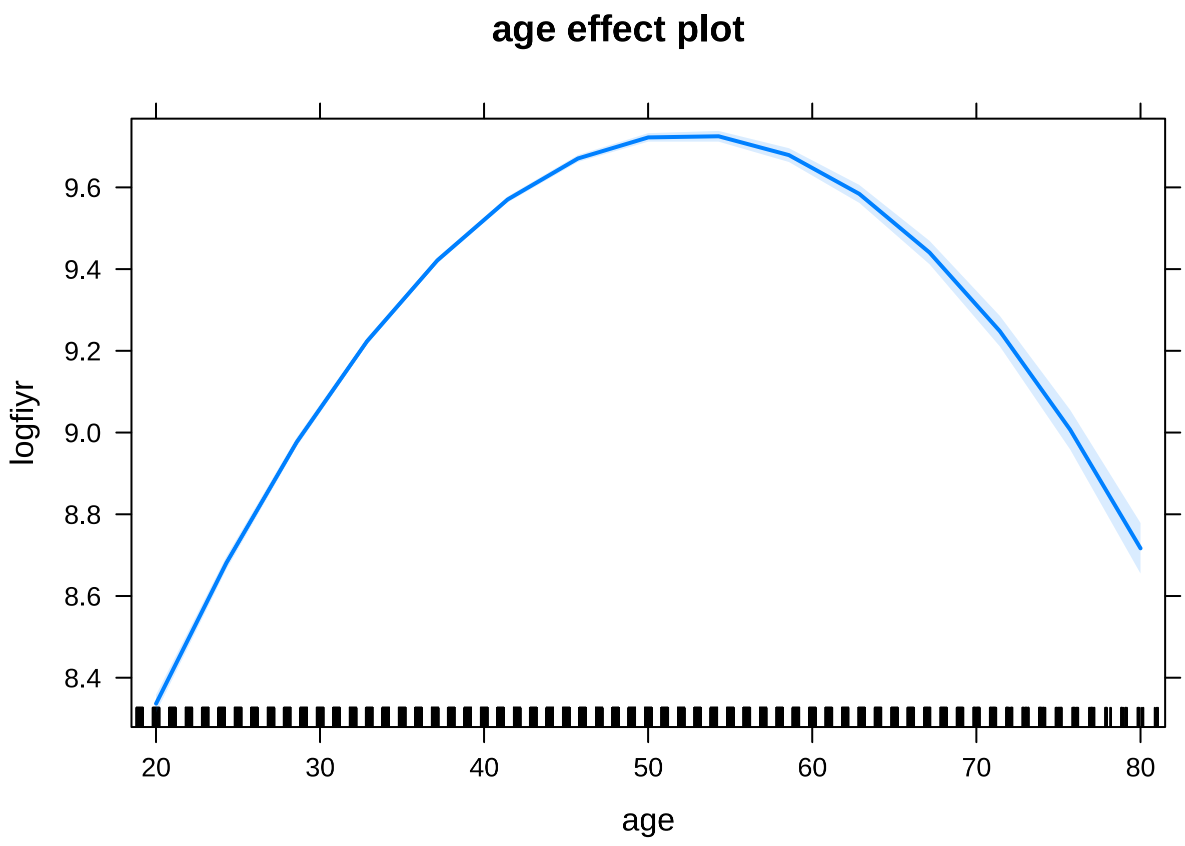
Next, using lmer a random intercepts model and one that explores whether sex differences in income have changed over time.

```
l1 <- lmer(logfiyr ~ sex + ft + age + I(age^2/1000) + (1 | pid), data = bhps_sub)
display(l1, detail = TRUE)
```

```
lmer(formula = logfiyr ~ sex + ft + age + I(age^2/1000) + (1 |
    pid), data = bhps_sub)
                       coef.est coef.se t value
(Intercept)              6.29     0.05  118.70
sexFemale               -0.42     0.02  -25.37
ftPart time: lt 30 hrs  -0.51     0.01  -51.56
age                      0.14     0.00   54.77
I(age^2/1000)           -1.33     0.03  -44.65

Error terms:
 Groups   Name        Std.Dev.
 pid      (Intercept) 0.60
 Residual             0.57
---
number of obs: 48647, groups: pid, 6882
AIC = 97307.6, DIC = 97213.3
deviance = 97253.5
```

```
plot(Effect("age", l1))
```



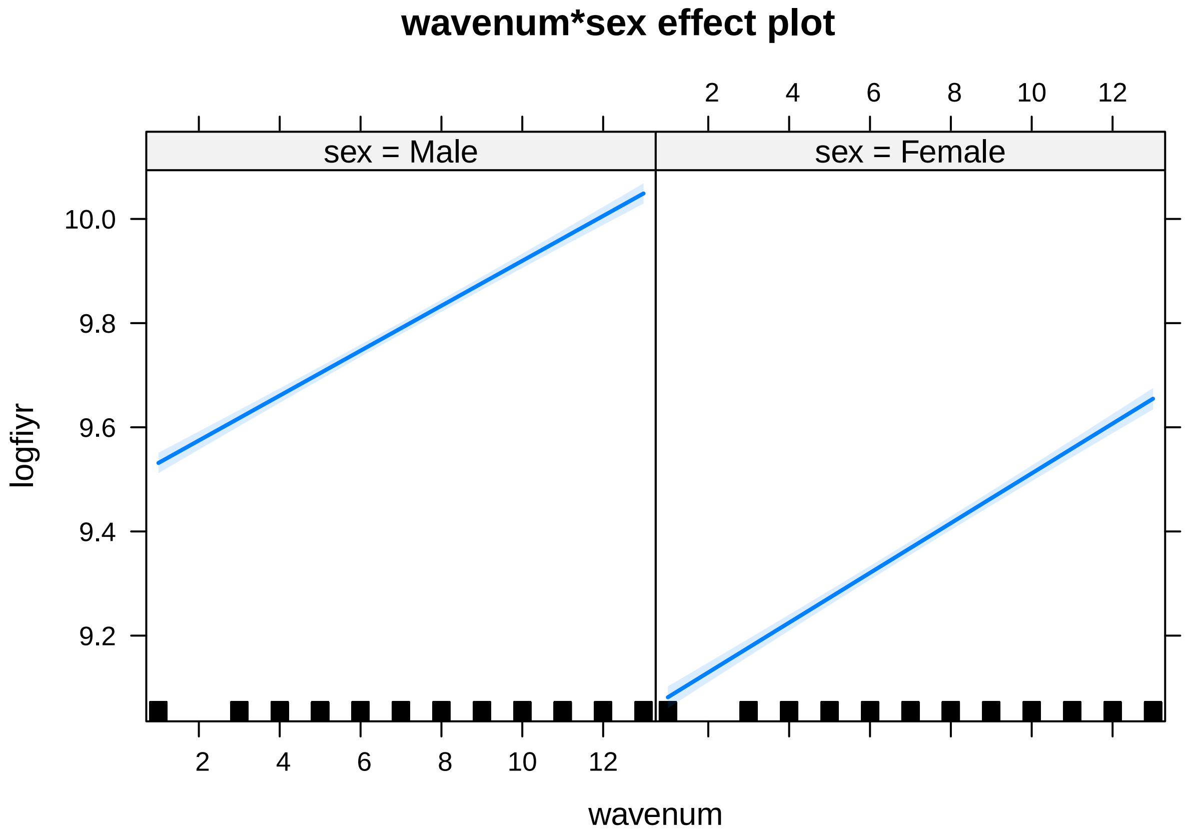
```
l2 <- lmer(logfiyr ~ sex * wavenum + ft + age + I(age^2/1000) + (1 + sex | pid),
    data = bhps_sub)

display(l2, digits = 4, detail = TRUE)
```

```
lmer(formula = logfiyr ~ sex * wavenum + ft + age + I(age^2/1000) +
    (1 + sex | pid), data = bhps_sub)
                        coef.est coef.se  t value
(Intercept)               7.1456   0.0536 133.3068
sexFemale                -0.4545   0.0178 -25.5338
wavenum                   0.0431   0.0012  35.7635
ftPart time: lt 30 hrs   -0.5177   0.0097 -53.2702
age                       0.1025   0.0025  40.2089
I(age^2/1000)            -1.1083   0.0290 -38.2173
sexFemale:wavenum         0.0046   0.0015   3.0402

Error terms:
 Groups   Name        Std.Dev. Corr
 pid      (Intercept) 0.5631
          sexFemale   0.4420   -0.4679
 Residual             0.5648
---
number of obs: 48647, groups: pid, 6882
AIC = 95132, DIC = 94982.1
deviance = 95046.0
```

```
plot(Effect(c("wavenum", "sex"), l2))
```



Change in employment

```
display(update(l2, . ~ . - ft + start_ft))
```

```
Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
$checkConv, : Model failed to converge with max|grad| = 0.00210821 (tol =
0.002, component 1)
```

```
lmer(formula = logfiyr ~ sex + wavenum + age + I(age^2/1000) +
    (1 + sex | pid) + start_ft + sex:wavenum, data = bhps_sub)
                  coef.est coef.se
(Intercept)        7.01     0.06
sexFemale         -0.70     0.02
wavenum            0.05     0.00
age                0.11     0.00
I(age^2/1000)     -1.24     0.03
start_ftYes       -0.07     0.02
sexFemale:wavenum  0.01     0.00

Error terms:
 Groups   Name        Std.Dev. Corr
 pid      (Intercept) 0.58
          sexFemale   0.50     -0.28
 Residual             0.55
---
number of obs: 44213, groups: pid, 6160
AIC = 86182.5, DIC = 86037.8
deviance = 86099.1
```

Homework

1. Use the dataset `Males` in the `plm` package. Explore the data.
2. The outcome variable of interest is `wage`.
3. Explore factors that influence wage, and in particular if there is evidence that married men earn more than single men. What problems might there be for drawing conclusions about this question based on these data?