Introduction to analysis of individual data using the apc.indiv functions in the packageapc

25 August 2020

Zoe Fannon Department of Economics, University of Oxford

Contents

1 Introduction

The purpose of this vignette is to demonstrate the use of some of the new commands available as part of the `apc.indiv` update to the R package `apc`.

This code is designed to allow the user to study the effects of age, period, and cohort on an outcome of interest. The age-period-cohort identification problem is avoided because the code uses the reparametrization approach developed in Kuang, Nielsen and Nielsen (2008). This approach does not attempt to separate the linear effects of age, period, and cohort, which are unidentified due to the well-known identification problem. Instead, the focus is on estimation of the non-linear effects. The non-linear effects that are identified are "double-differences" in each of age, period, and cohort. These "double-differences" are the accelerations in each of age, period, and cohort. By cumulating these accelerations a picture of the non-linear part of the relationship between age, period, or cohort and the outcome of interest can be constructed. Further details of the reparametrization approach and how the double-differences and cumulated double-differences should be interpreted are available in Nielsen (2015).

The new code allows for estimation of the reparametrized APC effects from the following:

- Gaussian models using repeated cross-section data
- Logistic models using repeated cross-section data
- Both of the above with survey weights
- Gaussian models using panel data (with POLS, random effects, and fixed effects options)
- All of the above with covariates included in the model

The tools build on several other packages. In particular `plm` (Croissant and Millo, 2008) and `survey` (Lumley, 2019) are used to perform the estimation for panel data and survey data respectively, while `lmtest` (Zeileis and Hothorn, 2002) and `car` (Fox and Weisberg, 2019) are used for testing restrictions. The aggregate-data functions from the package `apc` Nielsen (2015) were cannibalised extensively to produce the `apc.indiv` functions.

2 Repeated Cross Section

To illustrate the use of the code for repeated cross-section data, I use the `Wage` data from the `ISLR` package (James et al, 2017). This data records information about 3000 male workers in the Mid-Atlantic region of the US, and was manually assembled from the March 2011 supplement to the American Current Population Survey. I examine the age, period, and cohort effects on the log wage of these workers (a continuous outcome), and on the probability that they hold a job classified as "industrial" rather than "information" (a binary outcome). There is a concave non-linear relationship between age and the log wage, but no non-linear relationship with period or cohort. There is a sharp acceleration in the probability of holding an industrial job in 2008, followed by a compensating deceleration; this may indicate temporary layoffs in response to the financial crisis that are job class-specific. Note that this data does not contain weights and there is no evidence that the wage information has been corrected for inflation.

Data assessment and cleaning

I begin by examining the age-period-cohort structure of the data

```
>     library("plyr")
>     library("reshape")
>     library("ISLR")
>     data("Wage")
>     summary(Wage)
```

```
      year           age                     maritl           race     
 Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480  
 1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293  
 Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190  
 Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37  
 3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55                  
 Max.   :2009   Max.   :80.00                                          

               education                     region               jobclass   
 1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544  
 2. HS Grad        :971   1. New England       :   0   2. Information:1456  
 3. Some College   :650   3. East North Central:   0                        
 4. College Grad   :685   4. West North Central:   0                        
 5. Advanced Degree:426   5. South Atlantic    :   0                        
                          6. East South Central:   0                        
                          (Other)              :   0                        
            health      health_ins      logwage           wage       
 1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09  
 2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38  
                                     Median :4.653   Median :104.92  
                                     Mean   :4.654   Mean   :111.70  
                                     3rd Qu.:4.857   3rd Qu.:128.68  
                                     Max.   :5.763   Max.   :318.34  
```

```
>     AP_count <- count(Wage, c("age", "year"))
>     AP_show <- cast(AP_count, age~year)
>     AP_show[1:10,]
```

```
   age 2003 2004 2005 2006 2007 2008 2009
1   18    1    4   NA    4   NA    1    1
2   19    6   NA    2    2    1    1    2
```

```
3   20   4   2   NA   5   5   1   3
4   21   3   2    2   1   1   4   2
5   22   8   5    8   5   5   3   4
6   23   7   3   10   5   8  10   2
7   24   6   7    7   4   2   4   2
8   25  12   7   11   8   7   7   4
9   26  11   3    8   7   5   4   9
10  27   6   9    7   5   3  10  13
```

The output of the above is a long table, with a column for each of the seven periods in the data and a row for each of the 61 ages. Each cell shows the number of observations in the data for that age-period combination. The `apc.indiv` functions require a contiguous dataset, and so it is necessary to restrict the data by age and period so that no cells have 0 observations. In this case I will omit some of the youngest and oldest ages, which are sparsely observed.

```
>      Wage2 <- Wage[Wage$age >= 25 & Wage$age <= 55, ]
```

need to change the names

```
>      names(Wage2)[names(Wage2) %in% c("year","age")] <- c("period","age")
```

tidy some variables for the analysis

```
>      cohort <- Wage2$period - Wage2$age
>      indust_job <- ifelse(Wage2$jobclass=="1. Industrial", 1, 0)
>      hasdegree <- ifelse(Wage2$education
+               %in% c("4. College Grad", "5. Advanced Degree"), 1, 0)
>      married <- ifelse(Wage2$maritl == "2. Married", 1, 0)
>      Wage3 <- cbind(Wage2, cohort, indust_job, hasdegree, married)
```

In the above, I have restricted the data to those aged between 25 and 55. Note that I have also renamed some of the variables; the `apc.indiv` functions require that at least two of the variables `age`, `period`, and `cohort` are present in the data. I have also tidied some of the other variables that are of interest in the analysis, creating indicators for whether the job is industrial (as opposed to informational), whether the worker has a college degree, and whether the worker is married.

I will be interested in how the wage of the worker and the nature of their job (industrial or otherwise) is related to their age, cohort, and period of observation. Before performing a formal analysis of these relationships using the `apc.indiv` functions, I can use a visualisation to conduct a preliminary search for patterns in these variables along age, period, or cohort. This is done using `ggplot2` (Wickham, 2016).

```
>      library("ggplot2")
>      mean_logwage <- ddply(Wage3, .variables=c("period", "age"),
+      function(dfr, colnm){mean(dfr[, colnm])}, "logwage")
>      names(mean_logwage)[3] <- "Mean_logwage"
>     plot_mean_logwage <- ggplot(mean_logwage, aes(period, age)) +
+     theme_bw() +
+     xlab('\n Period') +
+     ylab('Age\n') +
+     geom_tile(aes(fill = Mean_logwage)) +
+     scale_fill_gradientn(colours=c("red", "blue"),
+     space = 'Lab', name="Mean logwage \n") +
+     scale_x_continuous(expand=c(0,0)) +
+     scale_y_continuous(expand=c(0,0)) +
+     theme(axis.text=element_text(size=18),
+     axis.title=element_text(size=24, face="bold"),
+     legend.title=element_text(size=20, face="bold"),
+     legend.key.size = unit(1, "cm"),
+     legend.text=element_text(size=18))

>     plot_mean_logwage
```

The output of the above code is seen in figure **??**. Each block shows the mean log wage among observations with that age-period combination in the data. The red colour corresponds to a lower mean log wage and the blue to a higher mean log wage. The concentration could be a combination of age and period effects: young people have lower wages, and in later years people have higher nominal wages (the data may not be adjusted for inflation).

We can use similar code to produce an analagous graph, showing the mean value of the indicator `indust_job` in each age-period cell. That mean indicates the proportion of people in that cell who have an industrial, rather than an information, role. This graph is not shown here; it has a similar pattern, with a higher probability of being in an industrial job concentrated among the young in the early 2000s.

Analysis of log wage

I now use the functions developed in `apc.indiv` to investigate the non-linear patterns in age, period, and cohort that can be identified from this data. These functions have a number of dependency packages which must be loaded if they are not already in the environment.

```
>     library("apc")
>     library("plyr")
>     library("lmtest")
>     library("car")
>     library("plm")
>     library("survey")
```

The first stage of the analysis is to estimate a table which can be used to determine how many of the elements of the age-period-cohort reparametrization are needed to describe the variation in this data. The relevant code is below: I specify the dataset, dependent variable, covariates I am including (in this case, I include the indicator for whether a person has a degree, as this is expected to influence their wage), the appropriate model family for this data, and which type of test I want to use. Here I chose a Wald test, compared against an F distribution; one could perform the Wald test using a Chi-squared distribution, or use a Likelihood Ratio test which must be compared with a Chi-squared distribution. I also choose to include a "TS" model in the table. This Time-Saturated (TS) model is a more general model than the reparametrized age-period-cohort model; it includes an indicator for each age-period combination present in the data. It nests the age-period-cohort model and therefore allows us to test whether the age-period-cohort model is sufficient to describe the variation in the data.

```
>     logwage_tab <- apc.indiv.model.table(Wage3, dep.var="logwage",
+     covariates="hasdegree", model.family="gaussian",
+     test="Wald", dist="F", TS=TRUE)
```

```
>     logwage_tab$table
```

```
    Wald (F) vs TS DF ( * , 2197) p-value Wald (F) vs APC DF ( * , 2342)
TS              NA              NA      NA                NA              NA
APC          1.052             145   0.323                NA              NA
AP           1.031             180   0.378             0.939              35
AC           1.063             150   0.292             1.356               5
PC           1.145             174   0.101             1.605              29
Ad           1.037             185   0.355             0.980              40
Pd           1.236             209   0.016             1.646              64
Cd           1.147             179   0.096             1.547              34
A            1.137             186   0.107             1.432              41
P            1.548             210   0.000             2.645              65
C            1.369             180   0.001             2.672              35
t            1.233             214   0.015             1.609              69
tA           1.301             215   0.003             1.810              70
tP           1.539             215   0.000             2.538              70
tC           1.385             215   0.000             2.068              70
1            1.613             216   0.000             2.749              71
    p-value      AIC      lik
TS       NA 1178.744 -370.372
APC      NA 1050.923 -451.461
AP    0.572 1014.561 -468.280
AC    0.238 1047.905 -454.952
PC    0.022 1040.461 -475.230
Ad    0.507 1010.992 -471.496
Pd    0.001 1029.152 -504.576
Cd    0.023 1036.545 -478.272
A     0.038 1028.702 -481.351
P     0.000 1091.992 -536.996
C     0.000 1075.497 -498.748
```

```
t      0.001 1024.753 -507.377
tA     0.000 1038.180 -515.090
tP     0.000 1087.515 -539.757
tC     0.000 1055.784 -523.892
1      0.000 1102.235 -548.117
```

The output of the above code is seen in table **??**. Look for the model which minimises the AIC; this is the model where the AIC takes the value 1010.99, i.e. the Ad model. We can also see that the p-values of the Wald tests of this model against the more general TS and APC models are quite large, indicating support for the reduction from those more general models. The Ad model, or age-drift model, includes double-differences in age only; the double-differences in period and cohort are constrained to zero. Further details of the APC sub-models are available in Nielsen (2015).

I now estimate the Ad model alone. I use a table to inspect the covariate coefficients, but the best way to examine the estimated time effects is by a visualisation. There are 30 ages in my dataset, which means 28 double-differences in age. Rather than looking at 28 estimates for double-differences, it is easier to understand their interpretation by plotting them.

```
>      logwage_ad <- apc.indiv.est.model(Wage3, dep.var = "logwage",
+      covariates="hasdegree",
+      model.family="gaussian",
+      model.design="Ad")
>      logwage_ad$coefficients.covariates

      Estimate Std. Error  t value      Pr(>|t|)
[1,] 0.2853405 0.01247669 22.86988 8.126105e-105

>      apc.plot.fit(logwage_ad, main.outer="")

WARNING apc.plot.fit: sdv large for plot 5 - possibly not plotted
```

As expected, the coefficient on having a degree is positive (0.285) and highly significant (p-value of $8.2e^{-105}$). The visual representation of the Ad part of the model is seen below. Subfigures (d) through (f) show the estimated linear plane, which combines the unidentified linear effects of age, period, and cohort. This is the “drift” part of the model. The first linear trend is plotted in the age dimension, while the second linear trend is plotted in the cohort dimension; respectively they combine the linear effects of age and period, and of cohort and period. The net effect of the three unidentifiable slopes then is that there is an increase in log wage with age and an increase in log wage with cohort, which may be used for forecasting purposes.

Subfigures (a) and (g) are of greater interest. Subfigure (a) shows the estimated double-differences in age, of which there are 28. Subfigure (g) shows the result of cumulating these to get a picture of the non-linear relationship between age and log-wage. We see that this relationship is concave, which would be consistent with acceleration in log wage up to the mid-30s and a plateauing thereafter.

This concavity is of interest because we can use it to evaluate consistency with theoretical models of the evolution of log wages over the life-cycle. For example, we could imagine a theory model which predicted that log wage is not only concave over the life cycle but is also quadratic. That the concavity is quadratic is a testable restriction in our APC model. The advantage to testing this quadratic hypothesis in this reparametrized APC model rather than another form of model is that this model has isolated the non-linear portion of age from the non-linear portion of cohort and period; therefore the test of the quadratic age effect is not contaminated by period or cohort effects.

We can perform this quadratic test as follows, using the `linearHypothesis` function from the package `car` (Fox and Weisberg, 2019).

```
>      allageDD <- rownames(logwage_ad$coefficients.canonical)[grep("DD_age",
+      rownames(logwage_ad$coefficients.canonical))]
>      ageDD1 <- allageDD[-1]
>      ageDD2 <- allageDD[-length(allageDD)]
>      quadratic_hyp <- paste(ageDD2, ageDD1, sep = " = ")
>      rm(list=ls(pattern="ageDD"))
>      linearHypothesis(logwage_ad$fit, quadratic_hyp, test="F")
```

```
Linear hypothesis test

Hypothesis:
DD_age_27 - DD_age_28 = 0
DD_age_28 - DD_age_29 = 0
DD_age_29 - DD_age_30 = 0
DD_age_30 - DD_age_31 = 0
DD_age_31 - DD_age_32 = 0
DD_age_32 - DD_age_33 = 0
DD_age_33 - DD_age_34 = 0
DD_age_34 - DD_age_35 = 0
DD_age_35 - DD_age_36 = 0
DD_age_36 - DD_age_37 = 0
DD_age_37 - DD_age_38 = 0
DD_age_38 - DD_age_39 = 0
DD_age_39 - DD_age_40 = 0
DD_age_40 - DD_age_41 = 0
DD_age_41 - DD_age_42 = 0
DD_age_42 - DD_age_43 = 0
DD_age_43 - DD_age_44 = 0
DD_age_44 - DD_age_45 = 0
DD_age_45 - DD_age_46 = 0
DD_age_46 - DD_age_47 = 0
DD_age_47 - DD_age_48 = 0
DD_age_48 - DD_age_49 = 0
DD_age_49 - DD_age_50 = 0
```

```
DD_age_50 - DD_age_51 = 0
DD_age_51 - DD_age_52 = 0
DD_age_52 - DD_age_53 = 0
DD_age_53 - DD_age_54 = 0
DD_age_54 - DD_age_55 = 0

Model 1: restricted model
Model 2: logwage ~ hasdegree + age_slope + cohort_slope + DD_age_27 +
    DD_age_28 + DD_age_29 + DD_age_30 + DD_age_31 + DD_age_32 +
    DD_age_33 + DD_age_34 + DD_age_35 + DD_age_36 + DD_age_37 +
    DD_age_38 + DD_age_39 + DD_age_40 + DD_age_41 + DD_age_42 +
    DD_age_43 + DD_age_44 + DD_age_45 + DD_age_46 + DD_age_47 +
    DD_age_48 + DD_age_49 + DD_age_50 + DD_age_51 + DD_age_52 +
    DD_age_53 + DD_age_54 + DD_age_55

  Res.Df Df      F Pr(>F)
1   2410
2   2382 28 1.2968 0.1368
```

Again, a Wald test is used, with comparison to an F distribution. The resulting test statistic of 1.297, with degrees of freedom $(28, 2382)$, has a p-value of 0.14. This indicates that the hypothesis of a quadratic relationship between the age of the worker and his log wage cannot be rejected.

Analysis of industrial job

The code can be used in a very similar way to investigate the relationship between a binary variable and age, period, and cohort. I illustrate this by building a model for whether or not the worker has an industrial job. Again, the analysis begins with a table comparing the time-saturated (TS) model, the full APC model, and submodels of the APC model.

```
>     indust_job_tab <- apc.indiv.model.table(Wage3, dep.var="indust_job",
+     covariates="hasdegree",
+     model.family="binomial",
+     test="LR", dist= "Chisq", TS=TRUE)

[1] "converged after 9 iterations"

>     indust_job_tab$table

    LR-test vs TS  df p-value LR-test vs APC df p-value      AIC Loglihood
TS             NA  NA      NA             NA NA      NA 3292.004 -1428.002
APC       169.954 145   0.077             NA NA      NA 3171.957 -1512.979
AP        235.345 180   0.004         65.391 35   0.001 3167.349 -1545.674
AC        179.286 150   0.052          9.332  5   0.097 3171.290 -1517.645
```

```
PC         206.755 174   0.045           36.801 29   0.151 3150.759 -1531.379
Ad         245.441 185   0.002           75.487 40   0.001 3167.444 -1550.722
Pd         271.016 209   0.002          101.062 64   0.002 3145.019 -1563.510
Cd         216.139 179   0.030           46.185 34   0.079 3150.142 -1536.071
A          245.464 186   0.002           75.510 41   0.001 3165.467 -1550.734
P          275.273 210   0.002          105.319 65   0.001 3147.276 -1565.638
C          216.345 180   0.033           46.391 35   0.094 3148.348 -1536.174
t          280.951 214   0.001          110.997 69   0.001 3144.954 -1568.477
tA         281.184 215   0.002          111.230 70   0.001 3143.188 -1568.594
tP         285.606 215   0.001          115.652 70   0.000 3147.610 -1570.805
tC         280.952 215   0.002          110.999 70   0.001 3142.956 -1568.478
1          285.784 216   0.001          115.830 71   0.001 3145.788 -1570.894
```

Note that here the model family is binomial, and the test used is a likelihood ratio test. The time-saturated model here is estimated by a custom Newton-Rhapson iteration procedure, so part of the output of this table is a report on the behaviour of that algorithm. If the print statement does not report convergence, the Newton Rhapson parameters should be modified using the option `NR.controls` until convergence is achieved - for example by increasing the number of iterations.

Looking at the estimated table, we see that the AIC is minimised towards the end of the table, by the tC model. However, the likelihood ratio tests comparing the tC model to the TS and APC models reject the restriction. Indeed most restrictions are rejected by the likelihood ratio test; the APC model itself is barely accepted as a restriction on the TS model. It is therefore difficult to select a model from this table. Ultimately I favour the PC model; it has one of the lower AIC values, is the most supported sub-model against the APC model, and is almost supported against the TS model. That said, this setting is one in which there is a strong argument that the APC model and its submodels do a poor job of capturing the time variation in the data, and some other reduction of the TS model should instead be used.

```
>      indust_job_pc <- apc.indiv.est.model(Wage3, dep.var="indust_job",
+      covariates="hasdegree",
+      model.family="binomial",
+      model.design="PC")
>      indust_job_pc$coefficients.covariates

       Estimate Std. Error   z value    Pr(>|z|)
[1,] -1.242142 0.09061988 -13.70717 9.19748e-43

>      apc.plot.fit(indust_job_pc)

WARNING apc.plot.fit: sdv large for plot 5 - possibly not plotted
```

Again, I directly estimate the preferred model using `apc.indiv.est.model` and inspect the estimated non-linearities in period and cohort using `apc.plot.fit`. There is a somewhat interesting pattern in the period effects, where there appears to be a substantial acceleration in the probability of having an industrial job in 2008; given that these are shipping workers, this may reflect a streamlining of operations during the financial crisis. However, there is no clear pattern in the cohort non-linearities. The effect of having a degree on the probability of having an industrial job is, unsurprisingly, significant and negative.

Extensions

The data I have been using does not include survey weights. However if they were present in the data, they could be quite easily added to all of the above analysis by simply specifying the name of the weight variable using the option `wt.var` in all of the above commands. It should be noted that since models incorporating survey weights are not estimated by maximum likelihood, the likelihood column is omitted from the data and one must use Wald tests rather than likelihood ratio tests. A psuedo-AIC is reported; see Lumley (2004, 2019) for details. Additionally, estimation of the time-saturated model has not yet been implemented for survey data, and so that will not be reported.

Sometimes the fact that the earliest and latest cohorts are only observed in one or two age-period cells can lead to instability in the estimates. This can be seen to some extent in the PC model for having an industrial job; the magnitude of the estimate for the earliest cohort is very large. This problem can be addressed by "censoring" those early and late cohorts out of the data, by first dropping them from the data using standard R techniques and then specifying the options `n.coh.excl.start` and `n.coh.excl.end`. The structure of this particular dataset is such that it can be displayed as a rectangle in age-period space. We say the data has an "age-period" format. In this case, it is the cohort double-differences where instability will appear and censoring should occur. However, not all data is of the "age-period" format. For example in the next section we will deal with data in "period-cohort" format; in that case, one would want to censor ages, using `n.age.excl.start` and `n.age.excl.end`. One might also have "age-cohort" data, in which case periods could be censored.

3 Panel data

To illustrate the use of the code for panel data, I use the `PSID7682` data from the `AER` package Kleiber and Zeileis (2008). This is an excerpt from the Panel Survey of Income Dynamics, covering 595 individuals over a seven-year period from 1976-1982 which has been used in economics textbooks such as Baltagi (2005) and Greene(2008). It is therefore an age-period dataset. Note that in this data the conflated variables are not age, period, and cohort, but rather years of work experience, period, and year of entering the workforce. There is an equivalent identification problem to the APC problem among these three variables, see for example Heckman and Robb (1985).

Data assessment and cleaning

I begin by inspecting the data. After initially displaying the data in an age-period format, it became clear that the period-cohort format was more appropriate.

```
>      library("plyr")
>      library("reshape")
>      library("AER")
>      data("PSID7682")
>      summary(PSID7682)
```

```
   experience        weeks       occupation  industry    south       smsa
 Min.   : 1.00   Min.   : 5.00   white:2036  no :2518   no :2956   no :1442
 1st Qu.:11.00   1st Qu.:46.00   blue :2129  yes:1647   yes:1209   yes:2723
 Median :18.00   Median :48.00
 Mean   :19.85   Mean   :46.81
 3rd Qu.:29.00   3rd Qu.:50.00
 Max.   :51.00   Max.   :52.00

 married       gender     union        education     ethnicity
 no : 773   male  :3696   no :2649   Min.   : 4.00   other:3864
 yes:3392   female: 469   yes:1516   1st Qu.:12.00   afam : 301
                                     Median :12.00
                                     Mean   :12.85
                                     3rd Qu.:16.00
                                     Max.   :17.00

      wage          year          id
 Min.   : 100.0   1976:595   1       :   7
 1st Qu.: 599.0   1977:595   2       :   7
 Median : 800.0   1978:595   3       :   7
 Mean   : 882.9   1979:595   4       :   7
 3rd Qu.:1046.0   1980:595   5       :   7
 Max.   :5100.0   1981:595   6       :   7
                  1982:595   (Other):4123
```

```
>      AP_count <- count(PSID7682, c("experience", "year"))
>      AP_show <- cast(AP_count, experience~year)
>      AP_show[1:10,]
```

```
   experience 1976 1977 1978 1979 1980 1981 1982
1           1    8   NA   NA   NA   NA   NA   NA
2           2   10    8   NA   NA   NA   NA   NA
3           3   35   10    8   NA   NA   NA   NA
4           4   19   35   10    8   NA   NA   NA
5           5   26   19   35   10    8   NA   NA
```

```
6           6  23  26  19  35  10   8  NA
7           7  30  23  26  19  35  10   8
8           8  25  30  23  26  19  35  10
9           9  15  25  30  23  26  19  35
10         10  34  15  25  30  23  26  19
```

The missing corners of the data show that this is actually cohort-period data (i.e. take a given set of people and follow them for X years, rather than observe people within an age group in a series of years).

```
>     period <- as.numeric(PSID7682$year) + 1975
>     entry <- period - PSID7682$experience
>     psid <- cbind(PSID7682, period, entry)
>     CP_count <- count(psid, c("entry", "year"))
>     CP_show <- cast(CP_count, entry˜year)
>     CP_show[1:10,]

   entry 1976 1977 1978 1979 1980 1981 1982
1   1931    1    1    1    1    1    1    1
2   1932    2    2    2    2    2    2    2
3   1936    6    6    6    6    6    6    6
4   1937    2    2    2    2    2    2    2
5   1938    3    3    3    3    3    3    3
6   1939    8    8    8    8    8    8    8
7   1940    6    6    6    6    6    6    6
8   1941   10   10   10   10   10   10   10
9   1942   13   13   13   13   13   13   13
10  1943    8    8    8    8    8    8    8
```

It is easily seen from CP_show that this is a balanced panel; the number of observations in a given cohort does not change over period. This makes it quite easy to see how we should restrict the data to ensure a sufficient number of observations in each cell. Again, I tidy some of the variables that will be used in the analysis, and rename the variables corresponding to age, period, and cohort.

```
>     psid2 <- psid[psid$entry >= 1939, ]
>     # which variables do we want to use?
>     logwage <- log(psid2$wage)
>     inunion <- ifelse(psid2$union == "yes", 1, 0)
>     insouth <- ifelse(psid2$south == "yes", 1, 0)
>     bluecollar <- ifelse(psid2$occupation == "blue", 1, 0)
>     # also education which is a continuous covariate
>
>     psid3 <- cbind(psid2, logwage, inunion, insouth, bluecollar)
>     names(psid3)[names(psid3) %in% c("experience","entry")] <- c("age","cohort")
```

It is important to visualise the data before estimating any models. This is done using ggplot2 in the same way as for repeated cross-sectional data. However, since this is period-cohort data, I plot cohort (year of entry) instead of age (experience) on the Y-axis.

```
>     library("ggplot2")
>     mean_logwage <- ddply(psid3, .variables=c("period", "cohort"),
+                       function(dfr, colnm){mean(dfr[, colnm])}, "logwage")
>     names(mean_logwage)[3] <- "Mean_logwage"
>     plot_mean_logwage <- ggplot(mean_logwage, aes(period, cohort)) +
+       theme_bw() +
+       xlab('\n Period') +
+       ylab('Entry \n') +
+       geom_tile(aes(fill = Mean_logwage)) +
+       scale_fill_gradientn(colours=c("red", "blue"),
+                            space = 'Lab', name="Mean logwage \n") +
+       scale_x_continuous(expand=c(0,0)) +
+       scale_y_continuous(expand=c(0,0)) +
+       theme(axis.text=element_text(size=18),
+             axis.title=element_text(size=24, face="bold"),
+             legend.title=element_text(size=20, face="bold"),
+             legend.key.size = unit(1, "cm"),
+             legend.text=element_text(size=18))

> plot_mean_logwage
```

I display the visualization in figure ??; note that in the labelling I have replaced "cohort" with "entry". There is a clear period effect; the colour becomes more blue towards the right of the graph, indicating higher wages in later years. There is also evidence of cohort effects, appearing as horizontal bands of colour. Those starting work around 1968, for instance, appear to have lower wages throughout their lives. That said, with a small panel we must be careful of confounding between cohort effects and individual fixed effects. Finally, age effects are evidenced by the predominance of red in the top-left corner of the graph; this is the area where individuals have the least work experience (those entering the workforce in the 1970s, observed in the 1970s), and we can unsurprisingly see that lack of experience means low wages.

To begin with, I consider a model with no covariates, just to get a sense of how the patterns seen in the graph above are reflected in a formal analysis. As was the case with repeated cross-sectional data, I begin with a table containing the full APC model and all submodels. Note that the time-saturated (TS) model is not currently implemented for panel data. Additionally, since the panel data model I consider (the random effects model) is not estimated by maximum likelihood, I lose the likelihood and AIC columns. Therefore model selection is by Wald test only.

```
>     library(apc)
>     panel_tab <- apc.indiv.model.table(psid3, dep.var="logwage",
+     model.family = "gaussian", test="Wald", dist="F",
+     plmmodel="random", id.var="id")

>     panel_tab$table

   Wald (F) vs APC DF( * , 3983) p-value
AP           1.974            35   0.001
AC           6.300             5   0.000
PC           2.593            41   0.000
Ad           2.510            40   0.000
Pd           2.388            76   0.000
Cd           3.278            46   0.000
A           48.106            41   0.000
P            2.701            77   0.000
t            2.790            81   0.000
tA          29.732            82   0.000
tP           3.079            82   0.000
```

It is clear from the table, seen in **??**, that none of the restrictions of the APC model pass muster. It is also worth noticing that some of the submodels seen in previous tables do not appear here. Those are: the C, tC, and 1 models. This is because random effects estimation requires at least one explanatory variable which changes over time within an individual, and these models do not satisfy this requirement.

The model selected by this analysis is, clearly, the APC model, and so I proceed to estimate and plot that using the standard tools.

```
>     panel_apc <- apc.indiv.est.model(psid3, dep.var="logwage",
+     model.family="gaussian",
+     plmmodel="random", id.var="id")
>     apc.plot.fit(panel_apc)

WARNING apc.plot.fit: sdv large for plot 5 - possibly not plotted
```

There is clear concavity in both age and period, while the non-linearity in cohort, despite being significant, lacks a clear pattern.

This model can also be estimated using fixed effects. This changes the set of models which are available, since the fixed effects are perfectly collinear with both the cohort double-differences and the combined slope that is estimated in the cohort dimension. The set of available models are as follows: FAP, FA, FP, Ft. These stand for “fixed effects with age and period non-linearities”, “fixed effects with age non-linearities”, “fixed effects with period non-linearities”, and “fixed effects with trend”. Note that FAP, FA, and FP all also contain the single linear trend that can be identified in these models, which is represented in the age dimension and combines the linear effects of age and period.

```
>     panel_tab_fe <- apc.indiv.model.table(psid3, dep.var="logwage",
+     covariates = c("inunion", "insouth",
+     "bluecollar"),
+     model.family = "gaussian", test="Wald", dist="F",
+     plmmodel="within", id.var="id")
>     panel_tab_fe$table
```

```
   Wald (F) vs FAP DF( * , 3436) p-value
FA           6.108             5       0
FP           2.557            41       0
Ft           3.215            46       0
```

Again, restrictions not accepted

```
>     panel_fap <- apc.indiv.est.model(psid3, dep.var="logwage",
+     covariates = c("inunion", "insouth",
+     "bluecollar", "education"),
+     model.family = "gaussian",
+     plmmodel="within", id.var="id",
+     model.design="FAP")
>     panel_fap$coefficients.covariates
```

```
                Estimate Std. Error     t-value   Pr(>|t|)
inunion      0.028285862 0.01513594  1.86878848 0.06173727
insouth      0.002419661 0.03404803  0.07106612 0.94334927
bluecollar -0.019586238 0.01400181 -1.39883641 0.16195230
```

```
>     apc.plot.fit(panel_fap)
```

The first step is to construct a table which compares all submodels to the most general model, which in the context of fixed effects is the FAP model. This table is not shown, as the conclusion is straightforward; even with the large set of covariates and the individual fixed effects, none of the submodel reductions is accepted. The FAP model is then estimated. The age and period non-linearities are largely unchanged from the random effects model, indicating that they are robust to the introduction of fixed effects and covariates. The covariate estimates are also shown. Most are not significant, which may be due to limited within-individual variation in those variables. Note that education was not included as a covariate: it has no within-individual variation and therefore its effect cannot be identified in a fixed effects model.

Extensions

It should be noted that at present panel data analysis only works for OLS models, so binary outcomes must be analysed in a linear probability framework. The time-saturated model is also not implemented for panel data. The censoring of cohorts, ages, or periods to improve the stability of estimates, described in section 2, is available for panel data.

In addition to the random and fixed effects models illustrated here, it is also possible to estimate panel data using pooled OLS. However, this is really no different to repeated cross section analysis.

4 References

Baltagi, B. H. (2005). *Econometric Analysis of Panel Data* (3rd ed.). Chichester, England: John Wiley & Sons.

Croissant, Y. and Millo, G. (2008) Panel data econometrics in R: The plm package. *Journal of Statistical Software* 27.

Greene, W. H. (2008). *Econometric Analysis.* Upper Saddle River, NJ: Pearson Prentice Hall.

Heckman, J. and Robb, R. (1985). *Using longitudinal data to estimate age, period and cohort effects in earnings equations*, pp. 137-150. New York, NY: Springer New York.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). ISLR: Data for an Introduction to Statistical Learning with Applications in R. R package version 1.2.

Kleiber, C. and Zeileis, A. (2008). *Applied Econometrics with R.* New York: Springer-Verlag.

Kuang, D., Nielsen, B. and Nielsen, J.P. (2008) Identification of the age-period-cohort model and the extended chain ladder model. *Biometrika* 95, 979-986. *Download*: Earlier version: http://www.nuffield.ox.ac.uk/economics/papers/2007/w5/KuangNielsenNielsen07.pdf.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9 (1), 1-19. R package verson 2.2.

Lumley, T. (2019) survey: analysis of complex survey samples. R package version 3.35-1.

Nielsen, B. (2015) apc: An R package for age-period-cohort analysis. *R Journal* 7, 52-64. *Download*: https://journal.r-project.org/archive/2015-2/nielsen.pdf.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer Verlag, New York.

Zeileis, A. and Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, 2 (3), 7-10.