

## LFPR Model (lfpr.prg)

The RATS program that performs the work when running the labor force participation model is named `lfpr.prg`. This program pulls data from two Fame databases: `cpscovar_rev.db` and `input.db`. All right-hand side (RHS) variables are pulled from `input.db`, while the labor force participation and population shares are pulled from `cpscovar_rev.db`.

The program `lfpr.prg` is not run directly but rather through “run” files. The “run” file defines and initializes all parameters and then sources the estimation code from `lfpr.prg`. The first set of parameters included in the “run” file control certain features and characteristics of the estimation process. These *estimation parameters* were established mainly for model development purposes and should remain essentially fixed. In contrast, *model parameters*, which specify the details of the model to fit, will typically be changed on a run-by-run basis. These parameters identify such things as the series to be included as independent variables and the estimation interval. Both types of parameters should be defined in a `.prg` file that can be submitted to rats from a command prompt (I refer to this file as a “run” file). The last line in the file should be a command to source the `lfpr.prg`. This has the effect of running the model with the parameters specified in the file.

## Instructions for Running the Model

*The zip file contains 23 different run files that were used for the paper. These files are labeled run1.prg – run23.prg (run5.prg is what is referred to as the “baseline” estimation throughout the paper – see the file `brook.txt` for a summary of the various files). To start a run, make sure that you are in the directory that contains the run file, the `lfpr.prg` file, and the databases `cpscovar_rev.db` and `input.db`. Open up the `lfpr.prg` program and modify line 7 :*

```
compute lfproutfile = "<output path>/lfpr_run"+%string(runno)+".db"
```

*Replace <output path> with the path of your current directory (unfortunately there’s no easy way of dealing with paths in RATs). Once this is done, simply type `rats run<run no>.prg`. **Note that the programs as written will currently only work in Unix/Linux. In addition, note that the data is read by the RATs program from a Fame database, meaning that your version of RATs must be able to read from such databases.***

*Once the run is complete, you can see some graphical summaries of the results by running the Fame procedure `see_results.pro`. From a Fame prompt, simply type the following:*

```
cloud "<path to pro files>/see_result.pro"  
$see_results <runno>, {m,q}, "<path to databases>","<path to pro files>"
```

*Where <path to pro files> is the path to the .pro files contained in the zip file, <path to databases> is the path to the input databases (`input.db` and `cpscovar_rev.db`) and the run output database (which should be the same if the instructions above were followed), and <runno> is the run number for which you would like to produce the summary. (Note that {m,q} is a namelist which specifies the frequency of the output you would like to view. If you ran only the monthly model or quarterly model, or would only like to see either the monthly or the quarterly if you ran both, replace this with either {m} or {q}. As written it will produce the summary for both monthly and quarterly estimates.) **Note that this procedure will only work correctly in Linux/Unix as it relies on some system commands.***

`see_results.pro` generates a lot of output both to the terminal and in the form of postscript files (for plots); it runs the programs `see_agg.pro` (to plot some aggregate results), `see_profiles.pro` (to plot the age and cohort effects), `see_contx.pro` (to plot simple contributions – see below for more details), `see_extra.pro`, `see_resids.pro` (various plots of model residuals), `see_ar.pro` (to plot autoregressive coefficients, if applicable), and `see_coefficients.pro` (to plot the estimated coefficients for RHS variables). Any of these programs can be run independently if desired (check the `see_results.pro` file for an example of how the program is typically run).

The contributions plots generated by the `see_contx.pro` procedure are not the same as those presented in some of the summary tables in the paper. This contributions produced by this program are obtained by simply keeping the RHS variable at its 1976Q1 level for non-cyclical variables or by setting the RHS variable to 0 for cyclical variables. **If you would like to duplicate the contributions as referenced in the paper, you should use the `contribs.pro` program. This program computes “chain-type” contributions by using the formula:**

$$C_{G,t'}^{v^i} = \sum_{a,s \in G} \sum_{t=t_0}^{t'} \text{MAVE} \left( \frac{\delta \text{lfpr}_{a,s,t}}{\delta v_{a,s}^i}, 2 \right) \beta_{a,s,v_i} \Delta v_{a,s}^i \times \text{MAVE} (S_{a,s,t}, 2)$$

Where  $C_{G,t'}^{v^i}$  is the cumulative contribution of variable  $v^i$  to the participation of group  $G$  (which may span any number of ages and include males or females or both). The subscripts  $a$ ,  $s$ , and  $t$  refer to an individual age (16-79), gender, and period respectively.  $\beta_{a,s,v_i}$  is the estimated coefficient for variable  $v^i$  for age  $a$  and gender  $s$ ,  $S_{a,s,t}$  is the share of the total 16+ population accounted for by individuals of age  $a$  and gender  $s$ , and  $\text{MAVE}(\dots, 2)$  represents a two-term moving average. Given the logit form of the regression, the derivative term becomes:

$$\frac{\delta \text{lfpr}_{a,s,t}}{\delta v_{a,s}^i} = \text{lfpr}_{a,s,t} (1 - \text{lfpr}_{a,s,t})$$

Because the contributions are computed on a period-by-period basis relative to the starting point,  $t_0$  (1976M1), the contributions for the cyclical variables may need to be normalized to set their zero level. Note however, that the net contribution between two periods can be computed by simply subtracting the two period's respective chain type contributions. These contributions are also additive across any number of variables. To run the `contribs.pro` procedure simply type the following from a Fame prompt:

```
cloud "<path to pro files>/contribs.pro"
contribs <runno>,*,<path to output database>
```

The computed  $C_{G,t'}^{v^i}$  will be displayed for each variable for the groups 16+, 25-54, 16-24, and 55+ (male, female and total). The groupings can be adjusted by modify the 2<sup>nd</sup> and 3<sup>rd</sup> default arguments to the procedure. In addition, the procedure will produce several postscript plots of the results.

## Program Details

If you are interested in creating your own specification (subject to the available data provided), the easiest way to create a “run” file is to simply copy an existing “run” file and modify the parameters as necessary. A brief description of the estimation and model parameters follows:

### Estimation Parameters

```
1. compute ae_poly_degree = -1
2. compute ce_poly_degree = -1
3. compute x_cohorts_beg   = 10
4. compute x_cohorts_end   = 10
5. compute reg_cohort_num  = 10
6. compute cohort_ex       = 10
7. compute logit           = 1
8. compute weighted        = 1
9. compute smoothed        = 1
10. compute errmod         = 0
11. compute stage2int       = 1
12. compute debug          = 0
```

1. **ae\_poly\_degree** is used to constrain the age effects to fit a polynomial of the specified degree. A value of -1 indicates that the effects are to be freely estimated (i.e. they are not constrained by any polynomial form, which is the default for all runs used for the paper).

2. **ce\_poly\_degree** is used to constrain the cohort effects to fit a polynomial of the specified degree. A value of -1 indicates that the effects are to be freely estimated (the default for all runs used for the paper).

3. **x\_cohorts\_beg** is used to specify the number of “older” cohorts that are to be excluded from the estimation. Data points that include terms for these cohorts will be excluded from the model estimation and the cohort effects will be estimated by the method selected using the **cohort\_ex** parameter (see below).

4. **x\_cohorts\_end** is used to specify the number of “younger” cohorts that are to be excluded from the estimation.

5. **reg\_cohort\_num** specifies the number of estimated cohorts to be used in extrapolating the cohorts excluded by the **x\_cohort\_beg** and **x\_cohort\_end** settings if the **cohort\_ex** parameter is set to either 0 or 2).

6. **cohort\_ex** is used to specify the method used to obtain the values of the cohort effects for excluded cohorts. A value of 0 indicates that the excluded cohorts effects should be derived using a simple linear extrapolation of the nearest **reg\_cohort\_num** of estimated cohort effects with the linear relationship being used to extrapolate the effects into the forecast period. A value of 1 indicates that all excluded cohort effects (including those in the forecast period) should be set to the value of the nearest estimated cohort effect. A value of 2, as with the 0 value, indicates that the nearest **reg\_cohort\_num** of estimated cohort effects should be used to linearly extrapolate the cohort effects. Unlike the case where the value is set to 1, the cohort effects for the forecast period are set to

equal the last observed value of the extrapolated cohort effects (i.e. the cohort effect for the last cohort that is covered by the estimation interval). The default method is 2.

7. **logit** is an indicator that specifies whether we would like to use a logit transformation on the data. 1 = "Yes", any other value indicates "No"

8. **weighted** is an indicator indicating whether a weighted regression is to be performed. A value of 1 indicates that the data should be weighted by:

$$w_{age}(t) = \frac{1}{n_{age}(t)} \frac{1}{lfpr_{age}(t)(1 - lfpr_{age}(t))}$$

This weighting is implemented by using this data as the "spread" factor in the RATS linreg procedure. The factor is intended to account for the heteroskedasticity arising from the differences in CPS sample size by age and time as well as the differing participation rates by age and time. In the above formula  $n_{age}(t)$  represents the number individuals aged "age" sampled at time  $t$  and  $lfpr_{age}(t)$  is the labor force participation rate for individuals aged "age" at time  $t$ .

9. **smoothed** is an indicator that determines the form of the cohort dummies used to capture the cohort effects. If set to 1, we refer to the cohort effects as smoothed as the cohort effect for a given birth year are spread across two years for any given age – and for any given age there are always two cohort effects in play). Any other value (i.e. 0) will not use the smoothed version (we refer to these as unsmoothed cohort effects). Equations for the cohort dummies in both cases follow (where  $p(t)$  is the period of time  $t$  and PPY is the number of periods per year):

$$K_{cohort\ year}^{unsmoothed}(age, t) = \begin{cases} 1 & \text{if } year(t) = cohort + age \\ 0 & \text{otherwise} \end{cases}$$

$$K_{cohort\ year}^{smoothed}(age, t) = \begin{cases} \frac{2 \times p(t) - 1}{2 \times PPY} & \text{if } year(t) = cohort + age \\ \frac{2(PPY - p(t)) - 1}{2 \times PPY} & \text{if } year(t) = cohort + age + 1 \\ 0 & \text{otherwise} \end{cases}$$

10. **errmod** select the type of error model used in the panel regression. A value of 0 indicates IID errors, a value of 1 indicates homogenous AR(1) errors, and a value of 2 heterogenous AR(1) errors.

11. **stage2int** indicates whether the second-stage regression (if applicable) should include an intercept term (1="Yes", any other value indicates "No").

12. **debug**. If set to 1 then some extra output will be displayed by the lfpr.prg code.

## Model Parameters

```
1. compute runno = 0
2. decl vect[int] allages(64); ewise allages(i) = 16+(i-1)
3. decl vect[int] educages(53); ewise educages(i) = 27+(i-1)
4. decl vect[int] ssages(18); ewise ssages(i) = 61+i
5. decl vect[int] ssdages(8); ewise ssdages(i) = 61+i
6. decl vect[int] lifeages(20); ewise lifeages(i) = 60+(i-1)
7. decl vect[int] teenages(4); ewise teenages(i) = 16+(i-1)

8. compute [vector[str]] rhsdesc = $
    || '+' Component of CBO Unemployment Gap", $
    || '-' Component of CBO Unemployment Gap", $
    ...
    ||;

9. compute [vec[str]] rhsvar = || 'cbogaplp', 'cbogapln', ... ||;
10. compute [vec[int]] rhsvartype = || 0 , 0 , ... ||;
11. compute [vec[int]] rhsvarsex = || 0 , 0 , ... ||;
12. compute [vec[int]] rhsvarnorm = || 1 , 1 , ... ||;
13. compute [vec[int]] rhsvarcyc = || 1 , 1 , ... ||;
14. compute [vec[int]] rhsvarstage = || 1 , 1 , ... ||;
15. compute [vec[vec[int]]] rhsages= ||allages ,allages , ... ||;

16. declare vec[vec[int]] rhslagl(%rows(rhsvar))
17. compute %do(i,1,%rows(rhsvar),rhslagl(i)=||0||)
18. compute rhslagl(1) = ||12,24,36||

* INFORMATION FOR FREQUENCY LOOP;
19. compute [vect[str]] fsv = || 'm', 'q' ||
20. compute [vect[int]] npersv = ||12, 4||
21. compute [vect[int]] syearv = ||1976,1976||
22. compute [vect[int]] sperv = ||1, 1||
23. compute [vect[int]] eyearv = ||2014,2014||
24. compute [vect[int]] eperv = ||6, 2||
25. compute [vect[int]] fyearv = ||2016,2016||
26. compute [vect[int]] fperv = ||12, 4||

27. source '/mcr/res_labor2/m1fxg00/labor_part/SSM/rats/lfpr_ssm.prg'
```

A brief description of each of the parameters (keyed to the highlighted line numbers) follows:

1. **runno** is an integer that is used to identify the output associated with the run. In particular, the run information is stored in a Fame database name <path>/lfpr\_run<runno>.db. If the output database already exists for the specified run number, it will be removed and recreated. It's a good idea to include this line near the top of a "run" file to make it easy to identify the run's run number at a later time.

2.-7. These integer vectors define various age ranges that will be used when defining the groups to which particular right-hand-side (RHS) variables apply. The first vector, named **allages**, has an entry for each of the ages modeled (16-79). The entries of the vector correspond to the actual ages that are to be included. You can add or remove as many vectors as needed so long as all the vectors referenced by **rhsages** (see below) are defined.

8. **rhsdesc** is a vector of strings whose elements provide a brief description or label for each of the RHS variables included in the model. **rhsdesc** must have the same number of elements as **rhsvar**.

9. **rhsvar** is a vector that provides the root names of the RHS variables in the model. The string represented in this array will be extended depending on the variable type.

10. For each RHS variable listed in **rhsvar**, the **rhsvar<sub>type</sub>** vector contains an integer that represents the variable type. Currently five variable types, represented by the integers 0-4, are supported. See the description of variable types below for information on how variable types are coded. Note that the variable type dictates the naming convention for the corresponding input series. **rhsvar<sub>type</sub>** must contain the same number of elements as **rhsvar**.

11. For each RHS variable listed in **rhsvar**, the **rhsvar<sub>sex</sub>** vector contains an integer that represents the genders for which the variable is applicable. 0 specifies that the variable is to be used in the regression for both men and women, 1 indicates that the variable will only be used in the regression for women, while 2 indicates that the variable is only to be used for men.

12. For each RHS variable in **rhsvar**, the **rhsvar<sub>norm</sub>** vector contains an indicator value indicating whether the corresponding RHS variable should be normalized before it is used in the estimation. Valid values are 0 and 1. **rhsvar<sub>norm</sub>** must contain the same number of elements as **rhsvar**.

13. For each RHS variable in **rhsvar**, the **rhsvar<sub>cyc</sub>** vector contains an indicator value indicating whether the corresponding RHS variable represents a cyclical control variable. The components of the fit associated with these variables are removed when determining the trend. **rhsvar<sub>cyc</sub>** must contain the same number of elements as **rhsvar**.

14. For each RHS variable listed in **rhsvar**, the **rhsvar<sub>stage</sub>** vector contains an integer that indicates whether the variable is to be included in the 1<sup>st</sup> or 2<sup>nd</sup> stage regression (set to 1 or 2 respectively). Currently the only variable that enters the second stage regression is the enrollment variable for teenagers.

15. **rhsages** is a vector of integer vectors whose component vectors represent the ages for which the corresponding RHS variable (in **rhsvar**) applies. These component vectors are those defined in lines 2-7. The number of **rhsages** component vectors must be the same as the number of elements in **rhsvar**.

16-17. These lines define a vector of integer vectors specifying the lags that are to be included in the model (if any) for the corresponding RHS variable. These lines simply initialize the vector and should be included as-is in all runs. To actually include lagged variables requires the addition of a line (comparable to line 18) for each RHS variable that is to be lagged.

18. For each variable for which lagged values are to be included in the model, a line similar to this line should be added. The index **i** to **rshlag1(i)** should match the index of the corresponding root name on the right-hand side of **rhsvar**. The right-hand side defines a vector of integers representing the lags (in months) to be included. As written, this line indicates that 1-year, 2-year, and 3-year lags of **cbo<sub>gaplp</sub>** should be included.

19-20. These lines define the estimation frequency or frequencies for the model. The vector **fsv** defines the strings that are expected to be found in (in the case of input series) or will be appended to (in the case of output series) the names of all series. The number of elements in **fsv** determines the number of frequencies for which the model will be run (in this case two, monthly and quarterly). The

vector in the next line, **npersv**, actually defines the frequency in periods-per-year format (12 for monthly and 4 for quarterly). Note that the strings in **fsv** do not in themselves specify a frequency. The strings could have any value, but they must coincide with whatever convention was used when defining the input series; for clarity it is best to keep these simple ('m' for monthly, 'q' for quarterly, and 'a' for annual).

21-26. For each estimation frequency defined in 18-19, these parameters (**syearv**, **sperv**, **eyearv**, **eperv**, **fyearv**, and **fperv**,) define the start and end periods for the estimation and the forecast respectively. The number of elements of each of these vectors must coincide with the number of elements of the vectors **fsv** and **npersv**.

## RHS Variable Description

As mentioned above, the model currently accepts five different variable types. The variable type specifies two distinct characteristics of the variable. The first dimension captures how the input itself varies by age-sex group. That is, is it a series that is the same for all age-sex groups? Is the value the same for both sexes but differs by age? Or – perhaps less usefully – is the input the same for all ages but differs by sex? The second dimension captured in the variable type is how the variable effect should be handled. Do we assume that the effect (i.e. coefficient) should be the same for all age-sex groups? Should it be the same across sex but differ by age? Or – again probably less usefully – should it be the same for all ages but differ by sex. The variable type also determines how the program expects to find the variable defined in input.db. For a RHS variable that is the same across age-sex groups only one input series is required whereas for series that differ by age-sex group the number of input series will be determined by the applicable number of ages for the variable as specified in the **rhsages** vector.

Below is a brief description of each of the variable types. In these description  $n_{in}$  represents the number of input series required,  $n_{age}$  represents the number of ages in **rhsage(i)** where **i** is the variable index in **rhsvar**, and  $n_{coef}$  indicates the number of coefficients estimated. Each description also specifies the name form of the series expected to be found in lfpr\_input\_test.db. In these descriptions <root name> is the corresponding entry in **rhsvar**, <f> indicates the data frequency, <age> is an integer, and <sex> is either "m" or "f". Finally, each description provides some examples of series that can be modeled as the given type. Note that for lagged variables, only the base series needs to be specified; all lags are computed within the RATS code.

0 = Input is the same for all age-sex groups. The coefficients vary by age-sex.

$$n_{in} = 1$$

$$n_{coef} = 2n_{age}$$

Name: <root\_name>\_<f>

Examples: Unemployment gap, UI Expenditures, Equity Wealth

1 = Input varies by age but not by sex. Coefficients vary by age-sex.

$$n_{in} = n_{age}$$

$$n_{coef} = 2n_{age}$$

Name: <root\_name>\_a<age>\_<f>

Examples: Social Security Payout, Current SS pay to SS at 70 pay ratio

2 = Input is the same for all age-sex groups. Coefficients vary by age *groups* but differ by sex.

$$n_{in} = 1$$

$$n_{coef} = 2n_{groups}$$

Name: <root\_name>\_<f>

Examples: none

3 = Input varies by age but not by sex. Coefficients are the same across all ages but differ by sex.

$$n_{in} = n_{age}$$

$$n_{coef} = 2$$

Name: <root\_name>\_a<age>\_<f>

Examples: Social Security variables (possibly)

4 = Input varies by age-sex. Coefficients vary by age-sex group.

$$n_{in} = 2n_{age}$$

$$n_{coef} = 2n_{age}$$

Name: <root\_name>\_a<age><sex>\_<f>

Examples: Life expectancy, educational attainment

5 = Input varies by sex but not by age. Coefficients vary by age-sex group.

$$n_{in} = 2$$

$$n_{coef} = 2n_{age}$$

Name: <root\_name>\_<sex>\_<f>

Examples: Wage ratio

6 = Input varies by age-sex. Coefficients vary by age *groups* but differ by sex.

$$n_{in} = 2n_{age}$$

$$n_{coef} = 2n_{groups}$$

Name: <root\_name>\_a<age><sex>\_<f>

Examples: none



## Output

lfpr.prg writes all of its output to the Fame database lfpr\_run<runno> .db. The output includes case series representing the run parameters as well as the output of the model run itself – the estimated coefficients, standard errors, fitted values, trend values, aggregated series, and some residuals. A selection of the series written to lfpr\_run<runno> .db is described below. In the tables that follow <I> is the index of the variable in **rhsvar**, <S> is either “m” for males or “f” for females, <A> an age integer between 16 and 79, <C> is the birth cohort year - 1000, <L> corresponds to a variable lag in months, <RHSVAR> is the root name of a right-hand side variable from **rhsvar**, and <F> is the frequency (“m” for monthly or “q” for quarterly).

### Model Parameters

<b>VARs</b>	String case series of RHS variable root names
<b>VARDESC</b>	String case series of RHS variable descriptions
<b>VARTYPE</b>	Numeric case series of RHS variable type
<b>VARNORM</b>	Numeric case series of RHS variable normalization indicators (0/1)
<b>VARCYC</b>	Numeric case series indicating if RHS variable is cyclical (0/1)
<b>VARLAGGED</b>	Numeric case series indicating if lagged RHS variables are included (0/1)
<b>V&lt;I&gt;_LAGS</b>	Numeric case series of lags for RHS variable indexed by <I> in R<RUNNO>_VARs

### Model Output

#### Cohort Effects

<b>CE&lt;S&gt;&lt;F&gt;</b>	Annual series of raw cohort effects (dates -> cohort birth year)
<b>CEAVG&lt;S&gt;&lt;F&gt;</b>	Annual series of cohort effects for average age effect
<b>CEMED&lt;S&gt;&lt;F&gt;</b>	Annual series of cohort effects for median age effect
<b>CEMIN&lt;S&gt;&lt;F&gt;</b>	Annual series of cohort effects for minimum age effect
<b>CEMAX&lt;S&gt;&lt;F&gt;</b>	Annual series of cohort effects for maximum age effect
<b>CE25&lt;S&gt;&lt;F&gt;</b>	Annual series of cohort effects for 25 <sup>th</sup> percentile age effect
<b>CE75&lt;S&gt;&lt;F&gt;</b>	Annual series of cohort effects for 75 <sup>th</sup> percentile age effect
<b>CE&lt;S&gt;&lt;C&gt;&lt;F&gt;</b>	Scalar raw cohort effect (same information as in R<R>_CE<S><F>)
<b>CESE&lt;S&gt;&lt;C&gt;&lt;F&gt;</b>	Scalar raw cohort effect standard error

#### Age effects

<b>AE&lt;S&gt;&lt;F&gt;</b>	Case series of raw age effects (cases 1-64 -> ages 16-79)
<b>AEAVG&lt;S&gt;&lt;F&gt;</b>	Case series of age profile for average cohort effect
<b>AEMED&lt;S&gt;&lt;F&gt;</b>	Case series of age profile for median cohort effect
<b>AEMIN&lt;S&gt;&lt;F&gt;</b>	Case series of age profile for minimum cohort effect
<b>AEMAX&lt;S&gt;&lt;F&gt;</b>	Case series of age profile for maximum cohort effect
<b>AE&lt;S&gt;&lt;A&gt;&lt;F&gt;</b>	Scalar raw age effect (same information as in R<R>_AE<S><F>)
<b>AESE&lt;S&gt;&lt;A&gt;&lt;F&gt;</b>	Scalar raw age effect standard errors

#### RHS Variables

<b>&lt;RHSVAR&gt;C&lt;L&gt;&lt;S&gt;&lt;F&gt;</b>	Case series of RHS variable coefficients (cases 1-64 -> ages 16-79)
<b>&lt;RHSVAR&gt;E&lt;L&gt;&lt;S&gt;&lt;F&gt;</b>	Case series of RHS variable coefficient standard errors