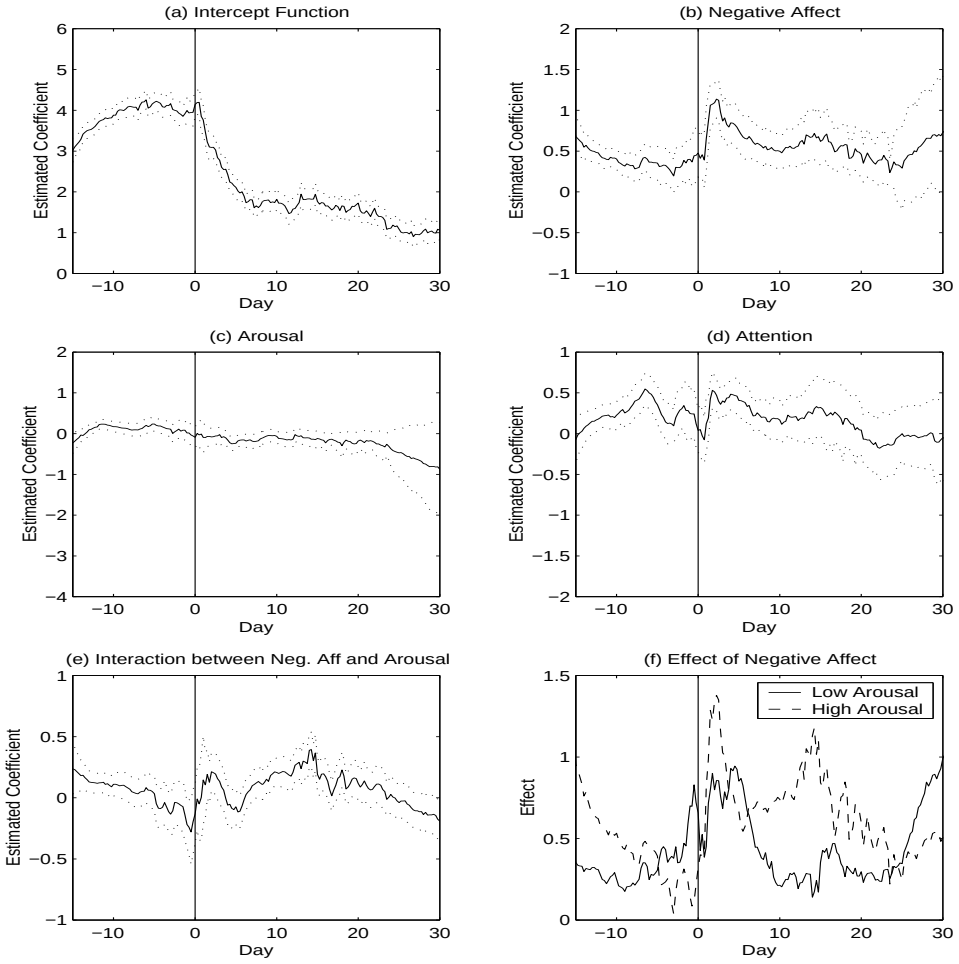
SAS Macro %FHLMLLR User's Guide (Version 1.0.0)

Runze Li and Xianming Tan
The Methodology Center
The Pennsylvania State University
University Park, PA 16802



Acknowledgments

The development of the SAS Macros %FHLMLLR was supported by National Institute on Drug Abuse, NIH, P50-DA10075 and National Institute on Drug Abuse, NIH, R21 DA024260.

Contents

1. Overview of This Macro

This user's guide describes the use of a SAS Macros: %FHLMLLR. This SAS Macro is a supplemental material of Li, Root and Shiffman (2006), in which the authors introduce functional multilevel modeling. This technique expands on the traditional linear mixed model by allowing coefficients to vary nonparametrically over time and introducing a local linear regression estimation procedure for the nonparametric coefficients. It is a powerful graphical tool that examines relationships in data that vary across time. The proposed model and estimation procedure are demonstrated using intensive longitudinal data, known as Ecological Momentary Assessment (EMA), on smoking cessation data. Functional multilevel models is also known as Functional Hierarchical Linear Model (FHLM). See Li, Root and Shiffman (2006) for more details.

This Macro is designed for users with SAS V 9.x (Windows Version). In addition, it also needs SAS/IML to conduct matrix manipulations, and PROC MIXED from SAS/STAT to estimate certain linear mixed effect models.

2. Functional Hierarchical Linear Model

Suppose that we observe intensive longitudinal data $\{(\mathbf{x}_{ij}, \mathbf{z}_{ij}, y_{ij}, t_{ij}), i = 1, 2, \ldots, n, j = 1, 2, \ldots, n_i\}$, where y_{ij} is the **response** of subject i measured at time t_{ij}, $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ and $\mathbf{z}_{ij} = (z_{ij1}, z_{ij2}, \ldots, z_{ijq})'$ are the corresponding p-dimensional and q-dimensional covariate vectors, respectively. It is allowed that $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ share some elements. Linear mixed effect model, also known as hierarchical linear model, is defined to be

$$y_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \boldsymbol{\gamma}_i{}'\mathbf{z}_{ij} + \epsilon_{ij}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of fixed effects, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2}, \ldots, \gamma_{iq})'$ denotes the random effects which is assumed to have normal distributions, say, $\gamma_{il} \sim N(0, \tau_l^2)$ for $l = 1, 2, \ldots, q$, and ϵ_{ij} is random error term with variance σ_2. In model (1), all coefficients in β are assumed to be constant over time, so are $\tau_l^2 (l = 1, 2, \ldots, q)$ and σ^2.

As a natural extension, functional hierarchical linear model (functional mixed effects

models, functional multilevel models) allows its effects changing over time without a pre-specified functional form.

$$y_{ij} = \boldsymbol{\beta}'(t_{ij}) \times \mathbf{x}_{ij} + \boldsymbol{\gamma}_i{}'(t_{ij}) \times \mathbf{z}_{ij} + \epsilon_{ij} \tag{2}$$

where $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \beta_2(\cdot), \ldots, \beta_p(\cdot))'$ consists of p unknown coefficient functions of time which are only required to be smooth, and $\boldsymbol{\gamma}_i(t_{ij}) = \Big(\gamma_{i1}(\cdot), \gamma_{i2}(\cdot), \ldots, \gamma_{iq}(\cdot)\Big)'$ consists of q random effect functions, with $\gamma_{il}(t) \sim N(0, \tau_l^2(t))(l = 1, 2, \ldots, q)$ and $\epsilon_{ij} \sim N(0, \sigma^2(t_{ij}))$.

The SAS Macro %FHLMLLR provides estimates of the functional effects $\boldsymbol{\beta}(\cdot)$ and variance functions $\tau_l^2(t)(l = 1, 2, \ldots, q)$ and $\sigma^2(t)$. It is important to note that conditional on t_{ij}, model (2) is a linear mixed effect model.

3. Estimation FHLM Using %FHLMLLR

In this section, we first illustrate the use of %FHLMLLR with an example, then describe the syntax of this Macro.

Example. Some observations of the SAS data set **electronic_diary.sas7bdat** (Appendix 2) are as follows:

idn	time	B_urge	int	NA	AR	AT	NA_AR
9	15.8153	3	1	-0.36178	0.87794	-0.60320	-0.31762
9	15.8563	2	1	-0.36178	0.87794	-0.60320	-0.31762
10	-15.0382	5	1	0.67822	-1.09206	-0.86320	-0.74066
10	-14.7090	10	1	2.41822	0.15794	-0.64320	0.38193
10	-14.6479	6	1	0.14822	-0.04206	-0.28320	-0.00623
10	-14.5625	5	1	-0.24178	-0.29206	-0.56320	0.07062
10	-14.4417	7	1	-0.13178	-0.36206	-0.55320	0.04771

where

idn: subject ID;

time: measurement time for each observation;

B_urge: the score of urge to smoke;

Int: the intercept term, all equal to 1;

NA: the centered score of negative affect;

AR: the centered score of arousal;

AT: the centered score of attention.

NA_AR: the multiplication of **NA** (negative affect) and **AR** (arousal).

We consider the following FHLM model to fit the data

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij}) \times x_{ij1} + \beta_2(t_{ij}) \times x_{ij2} + \beta_3(t_{ij}) \times x_{ij3} + \beta_4(t_{ij}) \times x_{ij1}x_{ij2} + \gamma_{1i}(t_{ij}) \times x_{ij1} + \gamma_{2i}(t_{ij}) \times x_{ij2} + \gamma_{3i}(t_{ij}) \times x_{ij3} + \epsilon_{ij}, \quad (3)$$

where

y_{ij}: the score of urge to smoke of the ith subject at time t_{ij};

x_{ij1}: the centered score of negative affect of the ith subject at time t_{ij};

x_{ij2} the centered score of arousal of the ith subject at time t_{ij};

x_{ij3}: the centered score of attention of the ith subject at time t_{ij}.

We apply the SAS Macro %FHLMLLR to estimate the functional effects and variance function as follows.

```
%FHLMLLR(
        mydata = electronic_diary,
        id = idn,
        time = time,
        dep = B_urge,
        tcov = int NA AR,
        random = NA AR AT,
        range = ,
        perct = 0.2,
        N = 100
        outfilename = c:/work/myplot.csv
);
```

The meanings of the first four parameters are self-explanatory. The 5^{th} parameter, **tcov**, lists those covariates which are assumed to have time-varying coefficients. The 6^{th} parameter, **random**, lists those covariates for random effects. The **range** parameter, which contains two numbers, say a and b, defines a time interval [a,b]. The default value for this parameter is the observed range of measurement times, i.e., a equals to the smallest measurement time and b the largest measurement time. The parameter **perct** determines how

many neighborhood observations will be included when conducting a local linear analysis for a given time point. The technical detail in Appendix 3 explains the meaning of the **perct** and how it affects the estimation. The last parameter **N** tells the Macro the number of grid points.

Roughly, the Macro first calculates N grid points: $t_i = a + \frac{(b-a)\times(i-1)}{(N-1)}, i = 1, 2, \ldots, N$ with a, b being determined by the parameter **range**; then, for each grid point t_i, **perct** percent of observations whose measurement times are closest to t_i will be included for this local linear analysis, which leads to the estimation of the values of the coefficient functions and variance functions at t_i. The last parameter specifies the path and name of a csv file, which contains the data for plotting coefficient curves and confidence bands.

Full Syntax Description

Table 1 lists all possible parameters of this Macro.

Parameter (=default)	Required	Description
mydata	Yes	The input data set; should have longitudinal data structure,
id	Yes	Subjects' identification variable
time	Yes	Measurement time variable
dep	Yes	Dependent variable
tcov	Yes	Covariates assumed to have time-varying coefficients. Notice that an all 1 variable, like **int** in the above example, should be included in this parameter if an intercept function is included in the model.
random	Yes	Covariates corresponding to those random effects in the model.
range	No	This parameter determines the time interval $[a, b]$ within which the local linear analysis will be conducted. In default, a is the smallest measurement time, and b is the largest measurement time among the pooled measurement times.
Perct(=0.3)	Yes	Determines how many observations will be included in a local linear analysis. Default value is 0.3.
N(=100)	No	This parameter determines how many local analysis will be conducted. Default value of this parameter is 100.
outfilename	No	Output file name. The Macro generates a csv file with its path and name specified by this parameter. This csv file contains the data for plotting the coefficient curves and their confidence bands for more specific use. By default, the csv file is saved in the temporary SAS working directory with name being "plot_data.csv". This parameter is available in v110 and after, but not in v100.

Appendix 1: A Simulated Example

In **simulated_data.sas7bdat**, we generated data for 50 subjects. For each subject, we schedule $J = 245$ measurement times evenly distributed over the interval of [0,1]. For subject i, two covariates, x_{ij1} and x_{ij2}, and an outcome variable, y_{ij}, are measured at

measurement times $t = \frac{j-0.5}{245}$, for $j = 1, \ldots, 245$. To generate x_{ij1} and x_{ij2}, we first generate two independent standard normal random numbers, say, z_{ij1} and z_{ij2}, then we let x_{ij1} equal to 1 if $z_{ij1} > 0$, and equal to 0 otherwise, and let $x_{ij2} = (z_{ij1} + z_{ij2})/\sqrt{2}$. In this way, x_{ij1} is a binary variable, and x_{ij2} follows a standard normal distribution. Furthermore, these two covariates are not independent but correlated, as in many practical situations.

We used the following model to generate y_{ij},

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij}) \times x_{ij1} + \beta_2(t_{ij}) \times x_{ij2} + \epsilon_{ij}, \tag{4}$$

where

$$\beta_0(t) = \exp(2t - 1),$$
$$\beta_1(t) = 8t(1 - t),$$
$$\beta_2(t) = 2\sin^2(2\pi t),$$

and the error term follows a AR(1) model, say,

$$\epsilon_{ij} = \rho \times \epsilon_{i,j-1} + e_{ij},$$

with e_{ij} being independent and identically distributed Gaussian random variables with variance σ^2. Specifically, in this example, we set $\rho = 0.3$ and $\sigma^2 = 1$.

To mimic a realistic situation, every subject may miss the actual schedule with 20% probability. This implies that, on average, we have $245 \times 0.8 = 196$ observations for each subject.Some observations in **simulated_data.sas7bdat** are listed as below:

subj	Y	t	X0	X1	X2
1	0.74397	0.00591	1	0	-1.71718
1	-0.68537	0.01378	1	0	-1.99814
1	0.20471	0.01772	1	1	0.59689
1	-0.63517	0.02559	1	1	-0.21331
1	2.08357	0.02953	1	1	1.10735
2	1.70600	0.00197	1	1	1.28186
2	0.49824	0.00591	1	1	1.55535

```
2      0.52604    0.00984     1     0    -0.84808
2      0.20461    0.01378     1     1     0.80863
2     -1.05567    0.02165     1     0     0.60380
```

We pretend that we do not know the exact form of the coefficient functions in (4), but try to estimate them based on observed data in **simulated_data.sas7bdat**. The SAS code is as follows:

```
%FHLMLLR(
        mydata = simulated_data,
        id = subj,  /* subjects' identification */
        time = t,   /* measurement time */
        dep = y,    /*  Y (outcome) */
        tcov = x0 x1 x2,    /* fixed effects */
        random = ,      /* no random effect in this model */
        range = 0 1,        /* time interval */
        perct = 0.2,    /*percentage of data points for a local model*/
        outfilename = c:/mysimulatedplot.csv  /* save data for figure as a file*/
);
```

The output of running the above code includes (1) three plots for the 3 estimated coefficient functions and 95% confidence band, and (2) a plot for the estimated variances function $\sigma^2(t)$ and 95% confidence band.

Appendix 2: An Electronic Diary Example

The data set **electronic_diary.sas7bdat** was inspired by a smoking cessation study (Shiffman et al., 2002). Specifically, we kept the original data structure but re-generated all observations. The data set consists of 149 smokers enrolled in the smoking cessation program which lasted about 45 days. Subjects were sampled at random, about 5 times a day, for assessment of affect and urge to smoke. Data on the intensity of subjects' urge to smoke was scored on a scale ranging from 0 to 10. For each subject, day 0 was the designated day for this subject to quit smoking.

Some observations in this data set are listed on Section 3.

The model we are interested in is:

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij}) \times x_{ij1} + \beta_2(t_{ij}) \times x_{ij2} + \beta_3(t_{ij}) \times x_{ij3} + \beta_4(t_{ij}) \times x_{ij1}x_{ij2} + \gamma_{1i}(t_{ij}) \times x_{ij1} + \gamma_{2i}(t_{ij}) \times x_{ij2} + \gamma_{3i}(t_{ij}) \times x_{ij3} + \epsilon_{ij}, \quad (5)$$

where

y_{ij}: the score of urge to smoke of the ith subject at time t_{ij};

x_{ij1}: the centered score of negative affect of the ith subject at time t_{ij};

x_{ij2} the centered score of arousal of the ith subject at time t_{ij};

x_{ij3}: the centered score of attention of the ith subject at time t_{ij}.

To estimate model (6) using **electronic_diary.sas7bdat**, we run the following SAS code:

```
%FHLMLLR(
        mydata = electronic_diary,
        id = idn,
        time = time,
        dep = B_urge,
        tcov = int NA AR,
        random = NA AR AT,
        range = ,
        perct = 0.2,
        N = 100,
        outfilename = c:/smokingplot.csv
);
```

The output, after running this code, contains

(1) five plots for the estimation and 95% confidence band of the 5 coefficient functions, $\beta_0(\cdot), \beta_1(\cdot), \beta_2(\cdot), \beta_3(\cdot), \beta_4(\cdot)$, respectively, and

(2) three plots for the estimation and 95% confidence band for the variances functions of $\tau_l^2(t)(l = 1, 2, 3)$, corresponding to the three random effect functions $(r_{1i}(t), r_{2i}(t), r_{3i}(t))$, respectively, (3) a plot for the estimation and 95% confidence band for the variance function $\tau_l^2(t)$ for the error term ϵ_{ij}.

Appendix 3: Technical Details

We briefly describe how local linear estimation procedure works to estimate FHLM. Interested readers could refer to Li, Root, & Shiffman (2006) for a more detailed description.

For the sake of clarity and simplicity, we consider the following simple model

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij}) \times x_{ij} + \gamma_{1i}(t_{ij}) \times x_{ij} \times x_{ij3} + \epsilon_{ij}, \tag{6}$$

where

$$\gamma_{1i}(t_{ij}) \sim N(0, \tau^2(t_{ij})) \; and \; \epsilon_{ij} \sim N(0, \sigma^2(t_{ij})).$$

We first consider estimating $(\beta_0(t_0), \beta_1(t_0), \tau^2(t_0), \sigma^2(t_0))$, i.e., the values of these functions at t_0. A key step is to locally and linearly approximate $\beta_0(t), \beta_1(t), \gamma_{1i}(t)$ in a neighborhood of t_0:

$$\beta_0(t) \approx \beta_{00} + \beta_{01} \times (t - t_0),$$

$$\beta_1(t) \approx \beta_{10} + \beta_{11} \times (t - t_0),$$

$$\gamma_{1i}(t) \approx \gamma_{1i0} + \gamma_{1i1} \times (t - t_0).$$

The accuracy of these approximations depends on the absolute distance from t to t_0: the smaller the absolute distance, the better the accuracy. To take this into account, we define a weight for each observation to gauge the contribution of this observation to the estimation of $(\beta_0(t_0), \beta_1(t_0), \tau^2(t_0), \sigma^2(t_0))$. Specifically, given a observation $\{(x_{ij}, y_{ij}, t_{ij}), i = 1, 2 \ldots, n, j = 1, 2, \ldots, n_i\}$, define its weight as $w_{ij} = h_0^{-1} K\{(t_{ij} - t_0)/h_0\}$, where

$$K(t) = \begin{cases} (3/4)(1 - t^2), & -1 \le t \le 1 \\ 0, & \text{otherwise} \end{cases}$$

h_0 is called bandwidth at t_0 which is determined by the Macro parameter **perct**. Given **perct**, to calculate h_0, we first sort the observations according to $d_{ij} = |t_{ij} - t_0|$, the absolute distance between t_0 and measurement time. Denote $d^{(l)}$ as the and l-th smallest distance, then $h_0 = d^{(p)}$ with p being the integer part of $(\sigma_i n_i) \times$ **perct**. It is worth to point out that the value of h_0 may vary with t_0.

Given w_{ij} for all observations, we define

$$
\begin{aligned}
&y_{ij}^* = \sqrt{w_{ij}} y_{ij},\ \epsilon_{ij}^* = \sqrt{w_{ij}} \epsilon_{ij}, \\
&x_{ij}^* = \sqrt{w_{ij}}[1, t_{ij} - t_0, x_{ij}, (t_{ij} - t_0)x_{ij}]',\ z_{ij}^* = \sqrt{w_{ij}}[x_{ij}, (t_{ij} - t_0)x_{ij}]', \\
&\beta^* = (\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11})',\ \gamma_i^* = (\gamma_{1i0}, \gamma_{1i1})'.
\end{aligned}
$$

We estimate β^* and γ_i^* by regarding them as regression coefficients in the following model

$$y_{ij}^* = (\beta^*)' x_{ij}^* + (\gamma_i^*)' z_{ij}^* + \epsilon_{ij}^*, \tag{7}$$

which can be viewed as a linear mixed model. Denote $\hat{\beta}^* = (\hat{\beta}_{00}, \hat{\beta}_{01}, \hat{\beta}_{10}, \hat{\beta}_{11})'$, $\hat{\gamma}_i^* = (\hat{\gamma}_{1i0}, \hat{\gamma}_{1i1})'$ as the estimates of β^* and γ_i^*, respectively, then

$$\hat{\beta}_0(t_0) = \hat{\beta}_{00},\ \hat{\beta}_1(t_0) = \hat{\beta}_{10},\ \textit{and}\ \hat{\gamma}_{1i}(t_0) = \hat{\gamma}_{1i0},$$

give the estimates for the three functions $\beta_0(t), \beta_1(t), \gamma_{1i}(t)$ at $t = t_0$, and the corresponding variance estimates from (7) given the estimation of $\tau^2(t), \sigma^2(t)$ at $t = t_0$.

Repeat the above procedure but replacing t_0 with $t_i (i = 1, \ldots, N)$, we obtain the estimate of these functions at the N grid points, and we can thus plot these functions by connecting these estimates.

Last but not the least, we can see that the parameter **perct** controls the percent of observations (with positive weights) be used to estimate (7), and the larger the **perct**, the more the observations will be included to estimate local linear mixed model like (7), hence, the larger the bandwidth h_0, and the smoother the estimated coefficient functions and variance functions.

Bibliography

[1] Li, R., Root, T., & Shiffman, S. (2006). A Local Linear Estimation Procedure for Functional Multilevel Modeling. *Models for intensive longitudinal data*. T. A. Walls and J. L. Schafer. New York, NY, US, Oxford University Press. : 63-83.

[2] Shiffman, S., Gwaltney, C.J., Balabanis, M.H., Liu, K.S., Paty, J.A., Kassel, J.D., Hickcox, M., Gnys, M. (2002). Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment. *Journal of Abnorm Psychology*. 111(4):531-45.