This product is part of the RAND Corporation technical report series. Reports may include research findings on a specific topic that is limited in scope; present discussions of the methodology employed in research; provide literature reviews, survey instruments, modeling exercises, guidelines for practitioners and research professionals, and supporting documentation; or deliver preliminary findings. All RAND reports undergo rigorous peer review to ensure that they meet high standards for research quality and objectivity.

TECHNICAL
REPORT

The Modified Kalman Filter Macro

User's Guide

Claude Messan Setodji, J. R. Lockwood,
Daniel F. McCaffrey, Marc N. Elliott,
John L. Adams

Sponsored by the Department of Health and Human Services

HEALTH

The research described in this report was sponsored by the Department of Health and Human Services and was conducted in RAND Health, a division of the RAND Corporation.

Published 2011 by the RAND Corporation
1776 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138
1200 South Hayes Street, Arlington, VA 22202-5050
4570 Fifth Avenue, Suite 600, Pittsburgh, PA 15213-2665
RAND URL: http://www.rand.org
To order RAND documents or to obtain additional information, contact
Distribution Services: Telephone: (310) 451-7002;
Fax: (310) 451-6915; Email: order@rand.org

Contents

1 Overview **5**
1.1 Macro Components and Implementation 7
1.2 Macro Features 8
1.3 Data Requirements 9

2 Getting Started **11**
2.1 Analyzing One Outcome Variable 12
2.2 Pooling Information from the Second Outcome 15

3 Syntax and Parameters **18**
3.1 Required Basic Parameters 18
3.2 Specification for Required Basic Parameters 18
3.3 Common Optional Parameters 20
3.4 Settings for Common Optional Parameters 20

4 Examples **27**
4.1 Example 1: Prevalence of Stroke in the United States in Different Racial/Ethnic Groups, (1997-2004 NHIS Data) 27
4.2 Disparities and Differences 31
4.3 Nonconform Data Error Messages 33
4.4 The Output Data File 33
4.5 Example 2: Subset Analysis of the Prevalence of a Disease Among Men and Women in Different Racial/Ethnic Groups 35
4.6 Example 3: Two outcomes, Prevalence of Hypertension and Diabetes in the United States in Different Racial/Ethnic Groups (1997-2004 NHIS Data). 37

5 Details and Theory **40**
5.1 Data Model 40
5.2 Prior Distributions for the Bayesian Models 41
5.2.1 Group Intercepts and Slopes 41
5.2.2 True State Deviations from Linear Trends 42
5.3 Implementation 42
5.4 Specification of Prior Distributions 43
5.5 Maximum Likelihood Estimation and Model-Averaging 44
5.6 Standard Errors of Zero 46

6 Appendix. Details for Linux Users **48**

7 Appendix. Advanced Macro Parameters **49**

7.1 Advanced Additional Parameters' Details 50

8 Acknowledgments: 52

9 Bibliography 53

1 Overview

There is much interest in using annual health outcomes' data to study racial/ethnic health disparities for both common racial/ethnic groups, such as non-Hispanic whites and blacks, and for rarer groups, such as American Indians/Alaska Natives (AI/AN), Asian and Hispanic subgroups. Examples of such health outcomes of interest include cancer, diabetes, hypertension, coronary heart disease (CHD), or average body mass index (BMI), which are frequently estimated from repeated cross-section samples of the US population through annual health surveys such as the National Health Interview Survey (NHIS). However, even large surveys like the NHIS typically provide only small annual samples of rarer subgroups, and the annual sample means can be very imprecise for these groups. For example, Table 1 provides 2004 NHIS estimates of the prevalence of stroke and diabetes for 11 racial/ethnic groups in the United States. The sampling errors are large for all but the most populous groups. For stroke, the relative standard error (standard error SE divided by the prevalence or mean) exceeds 0.20 for eight of the eleven groups and exceeds 0.30 for six of the groups. Typically, estimates with relative standard errors exceeding 0.3 are considered unstable (Klein et al., 2002) and are commonly suppressed from any inference/making decision because of their lack of reliability.

Table 1: Stroke and diabetes prevalence, standard errors and relative standard errors from the 2004 NHIS data and correlation between stroke and diabetes over time (1997-2004)

	2004 Prevalence of Stroke			2004 Prevalence of Diabetes			Correlation
Race/ethnicity	%	SE	Rel. SE	%	SE	Rel. SE	(1997-2004)
White	3.20	0.13	0.04	6.84	0.18	0.03	0.16
Black	3.01	0.28	0.09	10.34	0.49	0.05	0.40
AI/AN	3.77	1.44	0.38	16.04	2.76	0.17	0.49
Chinese	2.10	1.06	0.51	6.34	1.91	0.30	0.03
Filipino	3.09	1.40	0.45	8.28	1.98	0.24	-0.32
Asian Indian	0.70	0.69	1.00	10.47	2.43	0.23	-0.17
Puerto Rican	2.45	0.70	0.28	10.38	1.45	0.14	-0.05
Mexican	1.23	0.27	0.22	6.20	0.55	0.09	0.11
Cuban	3.09	1.54	0.50	12.18	2.24	0.18	-0.40
Other Hispanic	2.22	0.29	0.13	7.59	0.54	0.07	-0.04
All other	1.93	0.64	0.33	5.19	1.09	0.21	0.12

Note: Relative standard error is the standard error divided by the prevalence or mean.
The correlation is between the detrended diabetes and stroke prevalence.

As the magnitude of these standard errors is too large to meet the National Center for Health Statistics recommended standards for estimating health disparities (Klein et al., 2002), different methods have been proposed for better estimation of prevalence or means of interest (Lockwood et al., 2011). With population health outcomes generally evolving slowly over time, pooling data across years within groups provides an attractive means for improving the precision of the latest (current-year) annual estimates of disease prevalence and other health outcomes without increasing sample size. Co-morbid conditions can also be informative to disparity research in specific health outcomes. In a study of the disparity in diabetes between blacks and whites in the United States, Miller et al. (2004) reported that interventions addressing diabetes disparities should focus on managing co-morbidities, such as hypertension, shown to be related to the disparity. So, in the same manner, because of the clinical correlation between some health outcomes (e.g., diabetes and stroke), pooling data across outcomes and years simultaneously within groups can also help increase the precision of estimates. In the NHIS data in Table 1, even though from the same racial/ethnic groups, the relative standard errors of stroke in rarer racial/ethnic groups are above 0.30, making them unstable; the estimates for diabetes for the same groups have relative standard errors below 0.30. However, there is a significant correlation (time detrended) between stroke and diabetes in most racial/ethnic groups in the United States.

To improve precision, Elliott et al. (2009) developed a model called the Modified Kalman Filter (MKF), an extension of the Kalman filter estimation technique (Kalman, 1960) that assumes true health states in each racial/ethnic group evolve according to a group-specific linear trend and autoregressive (AR) deviations around that trend. They showed that the MKF is capable of improving the accuracy of health state estimates from such data as the NHIS. Lockwood et al. (2011) further extended the method to allow "borrowing information across groups" and Setodji et al. (2011) included information across correlated outcomes.

The MKF Procedure and MKF SAS© macro are designed to provide estimates of group means or prevalence rates from these different methods using data consisting of sample means and their standard errors from multiple time points within each of one or more groups. The MKF procedure pools data across time points within a group to improve the accuracy of the estimated mean for the final time point relative to the final period sample mean. When two outcomes are considered, where a correlation between those two outcomes can reasonably be assumed, this procedure also allows borrowing information from one outcome in the estimation of the other outcome. The sample means can be from simple random samples or complex survey designs.

The MKF macro models the sample means for any group as the unknown population mean plus an additive error term with variance given by user-supplied standard errors. The population mean is a function of a linear trend in time that describes the general progression of the outcome for the group and time period deviations from this trend. The goal of the software is to provide an accurate estimate of a population's means

(unknown trends plus unobserved deviations from the trend) given the model and the observed time period means. The macro

- estimates the model parameters;
- uses the estimated parameters to produce an optimal weighted average of linear trend and current and past years' estimates.

Using this approach, the MKF procedure can yield substantial gains in accuracy of estimates for small groups relative to a single time period sample mean (Elliott et al. 2009; Lockwood et al. 2011). The MKF macro produces estimates of a population means and the error in those estimates (i.e., an estimate of the root mean squared error, RMSE, which is analogous to the standard error of the sample mean). When dealing with a single outcome, Lockwood et al. (2009) derived a Bayesian implementation of the procedure that proved to be robust and provided an accurate assessment of the error in the predicted population means. The MKF macro offers users the choice of using the Bayesian implementation (the default when doing the estimation for a single outcome) or alternative (maximum likelihood based) estimation methods. When dealing with two outcomes, model-averaging based on two maximum likelihood estimation assumptions is the default. The model-averaging technique used in this macro can be applied both to the Bayesian as well as the maximum likelihood approach (with a single outcome), but as the Bayesian estimation uses a less stringent time trend assumption, the model-averaging approach is implemented in the maximum likelihood approach only to deal with maximum likelihood limitations of the specific slope assumption. With small sample sizes, flexibility in the time trend assumption was not warranted with maximum likelihood.

This software macro is design to be used by analysts for estimation and assumes familiarity with SAS© software. For better a understanding of the methods used in the macro, users are encourage to read the articles the macro is based on, including Elliott et al. (2009), Lockwood et al. (2011) and Setodji et al. (2011). Note that the macro is available for use in SAS and is not written for other commonly used statistical software, such as STATA©, SPSS©, or SUDAN©.

1.1 Macro Components and Implementation

The MKF macro software includes all the files for using the software under the Windows© or Linux© (Unix©) operating systems. The main User Guide provides details for using the software with the Windows© operating system. Details for using the software with the Linux© operating system are provided in Section 6.[1]

[1] The MKF macro will work with either the Windows© or Linux© (Unix©) operating systems without any changes to the SAS© macro code. However, the macro accesses an external executable file to

This macro requires two files that need to be saved together in a directory or folder chosen by the user. The files are

1. MKF_MACRO.SAS, the file containing the SAS© macro code that conducts the analysis;
2. kfwindows.exe , an external executable file accessed by SAS© to conduct statistical computation for Bayesian model estimation for a single outcome.

Users will need to refer to this directory via the software directory (software_dir) macro parameter (see details below) when implementing the MKF macro in SAS©. The macro creates temporary files in the system TEMP directory that are deleted after the macro is terminated.

1.2 Macro Features

The following are some of the basic features of the MKF macro:

- works with any type of group mean outcomes;
- works with any number of time periods greater than three;
- works with one or two outcomes:
  - for one outcome, information across time and groups is pooled;
  - for two outcomes, information across time, outcomes (correlation), and groups is pooled;
- allows the user to specify
  - the directory where the macro is stored;
  - group, time, outcome, and standard error variable names;
- allows users to choose
  - multiple subset or subgroup analyses;
  - either Bayesian or maximum likelihood estimation methods or both;
  - different specifications for time trends across groups (group-specific, common, or no time trends);
  - output specification;
- saves details of the statistical modeling to SAS© data sets that can be manipulated and saved by users.

conduct some of the statistical computations and the installation of this file is operating-system- dependent.

1.3 Data Requirements

The MKF estimation method uses group means and their associated standard errors. The group means are group (e.g., racial/ethnic) averages or prevalence that can be estimated from personal-level data over time, and the standard errors can also be estimated from personal-level data. The macro only allows for input of user's computed group means and standard errors, and as a first step before the use of the macro, using SAS or other statistical software, users should estimate these group means and standard errors, taking into account complex designs (e.g., sampling weights) when necessary before inputting them in the macro for estimation. The specific requirements of the macro are as follows:

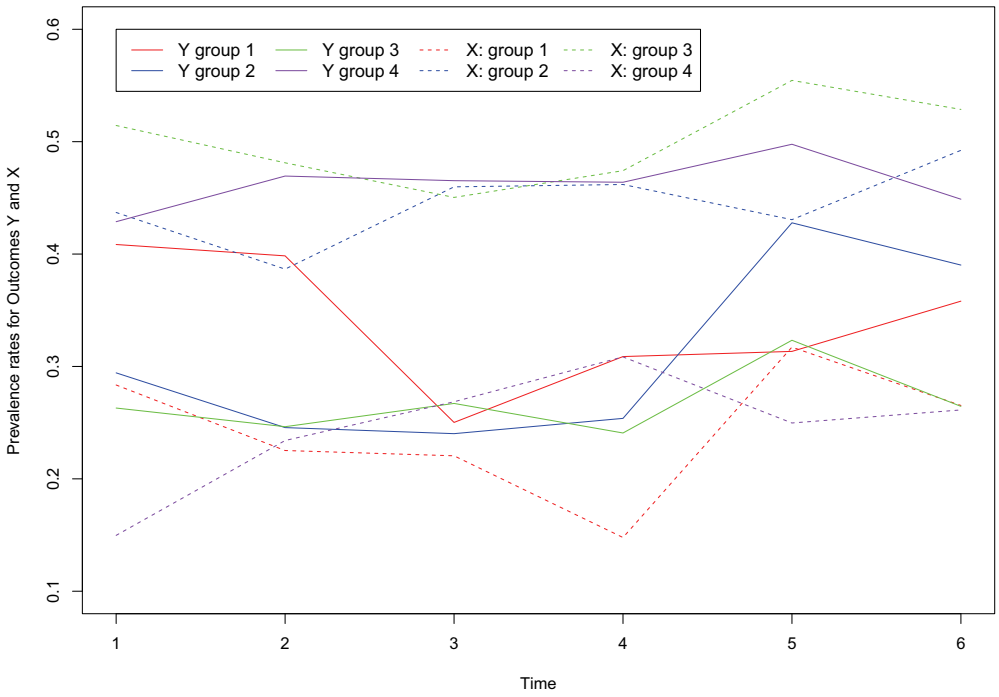
- The data should consist of one record for each of G groups measured at each of T time points for a total of $G \times T$ records.
- Every record must include a value for a group identifier variable to identify the G groups.
- The group identifier can be either a character or numeric variable.
- The time period must be numeric and equally spaced. For example, times could be $t_1 = 1$, $t_2 = 2$, etc., or $t_1 = 1998$, $t_2 = 2000$, $t_3 = 2002$, $t_4 = 2004$, etc., where the measurements are two years apart, but times could not be $t_1 = 1998$, $t_2 = 2000$, $t_3 = 2002$, $t_4 = 2003$, where from t_1 to t_2 or t_2 to t_3 there is a two-year span but between t_3 and t_4 there is only a one-year span.
- The outcome of interest Y_{gt} (and X_{gt} when dealing with two outcomes), $g = 1, \ldots, G$, $t = 1, \ldots, T$, can take any real value and will typically consist of group means or prevalence rates from samples of a population or subpopulation of interest.
- The outcome data must be complete, with no missing value for any group or time period.
- The data must contain standard errors ($SE_{Y_{gt}} \geq 0$) for each group estimated at each time point, with no missing values allowed, for all the outcomes of interest.
- An $SE_{Y_{gt}} = 0$ means that zero variance was observed in group g at time t, or that Y_{gt} for each member of the subpopulation g was the same. For each group, $SE_{Y_{gt}}$ (and $SE_{X_{gt}}$ if the interest is in two outcomes) must be greater than zero for at least one time period for the macro to produce population mean estimates. No variation within a group at a given time period might occur for rare diseases and small samples in which no cases in the sample have been observed with the disease.

Because missing data are not allowed in the macro, users who need to deal with missing values are encouraged to use missing data imputation techniques to fill in missing values before using the macro.

2 Getting Started

The following small example provides an introduction to the software. A researcher investigating the prevalence of a disease Y in four racial/ethnic groups (represented by numeric values $1,2,\ldots,4$) collected data on each group over six equally spaced time points. The researcher is interested in estimating the prevalence rate for each group for the most recent time point. The researcher actually collected information on a second disease X as well. The following plot presents the observations for the outcomes Y and X over the time $t = 1,2,\ldots,6$ for the different groups $1,2,3$ and 4.

Figure 1: Graph representation of the Y and X data over time



The following data step creates the data set for the analysis.

```
data diseases;
input group time Y sey X sex;
 datalines;
   1    1    0.4085    0.0328    0.2835    0.0420
   1    2    0.3984    0.0289    0.2252    0.0473
   1    3    0.2502    0.0324    0.2205    0.0478
   1    4    0.3088    0.0283    0.1478    0.0478
   1    5    0.3134    0.0345    0.3172    0.0495
   1    6    0.3581    0.0348    0.2653    0.0460
   2    1    0.2943    0.0278    0.4370    0.0491
   2    2    0.2456    0.0281    0.3866    0.0493
   2    3    0.2402    0.0272    0.4598    0.0432
   2    4    0.2538    0.0260    0.4619    0.0485
   2    5    0.4278    0.0245    0.4306    0.0499
   2    6    0.3902    0.0232    0.4923    0.0500
   3    1    0.2630    0.0211    0.5144    0.0494
   3    2    0.2464    0.0227    0.4812    0.0423
   3    3    0.2671    0.0216    0.4504    0.0465
   3    4    0.2408    0.0201    0.4742    0.0480
   3    5    0.3233    0.0249    0.5545    0.0472
   3    6    0.2645    0.0245    0.5287    0.0494
   4    1    0.4288    0.0213    0.1497    0.0416
   4    2    0.4694    0.0198    0.2341    0.0477
   4    3    0.4653    0.0208    0.2686    0.0498
   4    4    0.4639    0.0212    0.3085    0.0488
   4    5    0.4977    0.0214    0.2497    0.0405
   4    6    0.4488    0.0205    0.2613    0.0500
   ;
run;
```

The variable *GROUP* identifies the four racial/ethnic groups and *TIME* specifies the time point for each prevalence measurement. The diseases, prevalence rates and standard errors are defined as Y, X, and se_y, se_x, respectively.

2.1 Analyzing One Outcome Variable

When using only the outcome Y, the MKF macro call to produce the MKF estimates of the prevalence rates for the final time period is

```
%mkf(
      data= diseases ,        group= group ,
      time= time ,          outcome= Y ,
      se= se_y ,            software_dir= C: \SASAnalysis ,
      out= estimates
      )  ;
```

This macro specification identifies the data set (data= diseases), the group, time, outcome, and standard error variables (group= group, time= time, outcome= Y, se= se_y), the directory where the software is stored (software_dir= C: \SASAnalysis), and the name for a SAS© data set that will contain the MKF estimates and their RMSE estimates (out= estimates).[2] In SAS©, variables are not case-sensitive and that rule also applies to the macro.

The macro produces both a data set with the estimates and a printout of the results sent to the default SAS© output (e.g., the .lst file or the output window). For this example, the printout of the results is

```
                    MKF Full Bayesian Estimation for the outcome
                                          Y
      group  time  Estimation     Point       Std.   95% CI             Stdized  Relative
                     Type        Estimate    Error                      Diff     RMSE
#############################################################################################
      1      6     Sample        0.3581      0.0348  [ 0.2899,  0.4263] --       --
                   MKF estimate  0.3281      0.0284  [ 0.2725,  0.3836] -0.8634  0.8147
---------------------------------------------------------------------------------------------
      2      6     Sample        0.3902      0.0232  [ 0.3447,  0.4357] --       --
                   MKF estimate  0.3836      0.0208  [ 0.3429,  0.4244] -0.2834  0.8960
---------------------------------------------------------------------------------------------
      3      6     Sample        0.2645      0.0245  [ 0.2165,  0.3125] --       --
                   MKF estimate  0.2740      0.0206  [ 0.2336,  0.3144] 0.3887   0.8417
---------------------------------------------------------------------------------------------
      4      6     Sample        0.4488      0.0205  [ 0.4086,  0.4890] --       --
                   MKF estimate  0.4615      0.0180  [ 0.4262,  0.4969] 0.6217   0.8802
---------------------------------------------------------------------------------------------
```

As shown in the output, for group 1 at time 6, the sample mean prevalence estimate was $y_{16} = 0.3581$ with a standard error $se_{y16} = 0.0348$, while the MKF estimate was $\hat{y}_{16} = 0.3281$ with an RMSE of $\widehat{RMSE}_{16} = 0.0284$. A 95% confidence interval (CI) for the new point estimate can be computed as

$$95\%\, CI \quad = \quad \hat{y}_{16} \pm 1.96 \times \widehat{RMSE}_{16} \quad = \quad [0.2725, 0.3836].$$

[2] As discussed in Lockwood et al. (2009), the MKF procedure introduces bias into the estimates of the group means by pooling data across years. The variance of the error in predicting the group means equals the mean squared error, and the square root of the mean squared error is analogous to the standard error of the sample mean.

The standardized difference between the sample mean estimate and the MKF estimate,

$$\text{Stdized Diff } = \frac{\hat{y}_{16} - y_{16}}{se_{y16}} = -0.8634,$$

describes the difference between the MKF and the sample mean prevalence estimate relative to the sampling error in the sample mean. With the model providing better accuracy for the prevalence estimate, the relative root mean square error[3] equals

$$\text{Relative RMSE } = \frac{\widehat{RMSE}_{16}}{se_{y16}} = 0.8147.$$

After the macro completes calculation of the MKF estimates, it creates a SAS data set called ESTIMATES from the parameter estimates including the MKF estimates of the group prevalence rates for time 6 and their standard errors. In this example, the data set includes the sample mean prevalence estimates and standard errors (Y and se_y)and the MKF estimates and their RMSE for every group and each time point, Y_pred_B and Y_se_B. The output data set for this example is

[3] Because the sample mean is an unbiased estimate of the population mean, the standard error equals the RMSE of the sample mean.

group	time	Y	sey	y_pred_B	y_se_B
1	1	0.4085	0.0328	0.3773	0.0270
1	2	0.3984	0.0289	0.3694	0.0222
1	3	0.2502	0.0324	0.3155	0.0221
1	4	0.3088	0.0283	0.3249	0.0208
1	5	0.3134	0.0345	0.3204	0.0245
1	6	0.3581	0.0348	0.3281	0.0284
2	1	0.2943	0.0278	0.2631	0.0242
2	2	0.2456	0.0281	0.2596	0.0213
2	3	0.2402	0.0272	0.2755	0.0207
2	4	0.2538	0.0260	0.2954	0.0195
2	5	0.4278	0.0245	0.3833	0.0204
2	6	0.3902	0.0232	0.3836	0.0208
3	1	0.2630	0.0211	0.2574	0.0184
3	2	0.2464	0.0227	0.2519	0.0178
3	3	0.2671	0.0216	0.2654	0.0168
3	4	0.2408	0.0201	0.2545	0.0162
3	5	0.3233	0.0249	0.2954	0.0198
3	6	0.2645	0.0245	0.2740	0.0206
4	1	0.4288	0.0213	0.4388	0.0184
4	2	0.4694	0.0198	0.4622	0.0163
4	3	0.4653	0.0208	0.4629	0.0160
4	4	0.4639	0.0212	0.4648	0.0163
4	5	0.4977	0.0214	0.4837	0.0174
4	6	0.4488	0.0205	0.4615	0.0180

2.2 Pooling Information from the Second Outcome

For combining information from the second outcome X, making use of the correlation between Y and X, the MKF macro call to produce the MKF estimates of the prevalence rates is:

```
%mkf(
      data= diseases ,        group= group ,        time= time,
      outcome= Y ,      se= sey ,   outcome2= X ,      se2= sex,
       software_dir= C: \SASAnalysis ,      out= estimates2
      )  ;
```

In addition to the parameters specified in the one outcome setup, two parameters are now added: The specifications outcome2=X and se2=se_x refer to X and se_x as the additional outcome and its standard error, respectively, and this new outcome has the same requirements as the outcome defined previously.

The printout of the results of this run will be as follows:

```
                MKF group slope MLE Model Averaging Estimation for the outcome
                                                 Y and X
        group  time  Estimation     Point       Std.   95% CI             Stdized  Relative
                       Type        Estimate    Error                      Diff     RMSE
#############################################################################################
        1      6          Y Estimation:
                      Sample        0.3581      0.1065  [ 0.1494,  0.5668] --       --
                      MKF estimate  0.3272      0.0877  [ 0.1553,  0.4991] -0.2903  0.8234
                          X Estimation:
                      Sample        0.2653      0.0334  [ 0.1998,  0.3308] --       --
                      MKF estimate  0.2577      0.0252  [ 0.2083,  0.3071] -0.2277  0.7539
---------------------------------------------------------------------------------------------
        2      6          Y Estimation:
                      Sample        0.3902      0.1138  [ 0.1671,  0.6133] --       --
                      MKF estimate  0.3671      0.0918  [ 0.1871,  0.5471] -0.2032  0.8069
                          X Estimation:
                      Sample        0.4923      0.0357  [ 0.4223,  0.5623] --       --
                      MKF estimate  0.4758      0.0264  [ 0.4240,  0.5275] -0.4627  0.7386
---------------------------------------------------------------------------------------------
        3      6          Y Estimation:
                      Sample        0.2645      0.1126  [ 0.0438,  0.4852] --       --
                      MKF estimate  0.2836      0.0901  [ 0.1071,  0.4601] 0.1696   0.7996
                          X Estimation:
                      Sample        0.5287      0.0354  [ 0.4594,  0.5980] --       --
                      MKF estimate  0.5248      0.0259  [ 0.4741,  0.5755] -0.1098  0.7321
---------------------------------------------------------------------------------------------
        4      6          Y Estimation:
                      Sample        0.4488      0.1136  [ 0.2262,  0.6714] --       --
                      MKF estimate  0.4752      0.0863  [ 0.3061,  0.6444] 0.2327   0.7600
                          X Estimation:
                      Sample        0.2613      0.0356  [ 0.1914,  0.3312] --       --
                      MKF estimate  0.2798      0.0248  [ 0.2312,  0.3284] 0.5188   0.6961
---------------------------------------------------------------------------------------------
```

Then the output estimate2 will be as follows (only the first eight lines are printed here):

```
                                                    Y_pred_              X_pred_
group   Time     Y        sey      X        sex       MA       Y_se_MA     MA       X_se_MA

  1       1    0.4085   0.0328   0.2835   0.0420    0.36416   0.082763   0.23509   0.023782
  1       2    0.3984   0.0289   0.2252   0.0473    0.34966   0.066449   0.23752   0.019118
  1       3    0.2502   0.0324   0.2205   0.0478    0.33788   0.056171   0.24069   0.016221
  1       4    0.3088   0.0283   0.1478   0.0478    0.32455   0.057343   0.24343   0.016552
  1       5    0.3134   0.0345   0.3172   0.0495    0.33723   0.069827   0.25388   0.020084
  1       6    0.3581   0.0348   0.2653   0.0460    0.32718   0.087700   0.25769   0.025204
  2       1    0.2943   0.0278   0.4370   0.0491    0.25933   0.090204   0.41898   0.025925
  2       2    0.2456   0.0281   0.3866   0.0493    0.27171   0.070306   0.42762   0.020216
  .....................................................................................
 Y_pred_              X_pred_              Y_pred_              X_pred_
    G        Y_se_G      G        X_se_G      1        Y_se_1      1        X_se_1

 0.37846   0.086733   0.24238   0.023507   0.34445   0.076960   0.22503   0.024157
 0.35750   0.069200   0.24181   0.018755   0.33885   0.062462   0.23161   0.019606
 0.34161   0.057326   0.24262   0.015537   0.33275   0.054540   0.23804   0.017120
 0.32356   0.058680   0.24284   0.015903   0.32591   0.055449   0.24423   0.017405
 0.32329   0.072814   0.24788   0.019734   0.35644   0.065489   0.26216   0.020557
 0.30355   0.091845   0.24764   0.024892   0.35974   0.081645   0.27153   0.025628
 0.23150   0.094457   0.41695   0.025600   0.29767   0.083994   0.42178   0.026365
 0.25659   0.073422   0.42698   0.019899   0.29254   0.065771   0.42851   0.020645
  .....................................................................................
```

The variable configuration in this output is similar to the previous one where the sample mean prevalence estimates and standard errors of the two outcomes (Y, se_y, X, and se_x) are reported in addition to the group and time variables and the MKF estimates. The default method for the two-outcome analysis is maximum likelihood estimation (MLE) Model-Averaging Estimation, which does model-averaging of a model assuming different slopes for each group and a second model with same slope for all the groups. Therefore, the default output reports the Model Averaging Estimation for the two outcomes (Y_pred_MA and X_pred_MA) as well as their RMSE (Y_se_MA and X_se_MA) for every group and each time point. Then the model with group slope estimations ($Y\_pred\_G, Y\_se\_G, X\_pred\_G$, and X_se_G) and the ones with single slope estimations ($Y\_pred\_1, Y\_se\_1, X\_pred\_1$, and X_se_1) are also reported.

3 Syntax and Parameters

This section provides the basic syntax and macro parameter names and default values for calling the MKF macro. It lists the basic required parameters that must be specified by the user for every run of the macro and common optional parameters that control the structure of the input data set, the estimation method, the presentation of the results, and the variables in the output data set. Additional parameters for controlling statistical computations are provided in Section 7.

3.1 Required Basic Parameters

The following parameters are required in the MKF macro. By setting only these parameter values, users can run the MKF estimation procedure for one outcome using default settings for modeling choices and computational methods and produce the standard printout.

data = < SAS© data set name >,

group = <variable name >,

time = <variable name >,

outcome = <variable name>,

se = <variable name>,

software_dir = <directory name>.

If, in addition, one has two outcomes and decides to take advantage of the correlation between the two outcomes, the following default parameters should be added:

outcome2 = <variable name>,

se2 = <variable name>.

3.2 Specification for Required Basic Parameters

data = SAS© data set name

specifies the valid SAS© data set name for the data set containing the sample means and their standard errors for every group and time period. It must also contain the group and the period. It can be a permanent or temporary data set. The standard format will be one record per group per period. The record will have at least four variables: the group, the period, the outcome, and the standard error (records can contain other variables but they will be ignored by the macro; they

will be carried through the macro without any impact on the analysis). When two outcomes are used, both outcomes will need to be in separate variables in the data and will be used appropriately. The data do not need to be sorted.

group = variable name
the name of the variable defining the groups or subpopulations in the data. The group values can be numeric or characters, but missing values are not allowed. Although the data do not need to be sorted by group , the output will be sorted by the order of first appearance of the values of the group and within group, from the latest to the earliest period in the input data set. If a specific ordering of the output is desired, the user should input or sort the data accordingly.

time = variable name
the *numeric* time variable as specified in the data. Missing time points are not allowed. All groups must have the same number of time points and all time points must be equally spaced.

outcome = variable name
the *numeric* outcome variable name as specified in the input data set. Missing values are not allowed.

se = variable name
the *numeric* outcome variable name for group, time-period specific standard error of the sample mean outcome. (Users are reminded to check that data include the standard error and not their squares or the standard deviations.) Missing values are not allowed, and any standard error equal to zero will be imputed using the average of non-zero standard errors within the group. If all standard errors within a group are zero, the MKF estimation will not be conducted.

software_dir = directory name
specifies the directory where the macro is saved along with the executable files used to conduct the statistical computations.

outcome2 = variable name
the *numeric* second outcome variable name as specified in the input data set, if taking advantage of correlation between outcomes. Missing values are not allowed.

se2 = variable name
the *numeric* outcome variable name for group, time-period specific standard error of the second sample mean outcome. (Users are reminded again to check that data include the standard error and not their squares or the standard deviations.) Missing values are not allowed and any standard error equal to zero will be imputed

using the average of non-zero standard errors within the group. If all standard errors within a group are zero, the MKF estimation will not be conducted.

3.3 Common Optional Parameters

The data, group, time, outcome, se, and software_dir parameters are required (and outcome2 , se2 when using two outcomes) while all other parameters are optional and have default values that will be used if not altered by the user. The optional parameters and their valid values are described below.

out = <optional SAS© data set name, default=param >,

system = <optional operating system name, default=windows>,

by = <optional variable name, default= (empty)>,

comparedto = <optional group name, default= (empty)>,

comparedata = <optional SAS© data set name, default=&out._diff >,

slopes = <optional keyword, default= (empty)>,

bayesmodel = <optional keyword, default=full >,

modelprint = <optional keyword, default=no>,

finalprint = <optional keyword, default=yes>,

xtrakeep = <optional variable names, default= (empty)>,

pdigit = <optional integer, default=4 >.

3.4 Settings for Common Optional Parameters

out = optional SAS© data set name

specifies the name of the data set where the model estimates are stored. The data set name can be any valid SAS© data set name and the data set can be permanent or temporary. The default value is param if no value is provided by the user. The output data set will include the user-specified values of the outcomes, the SEs, the time periods, and the groups (along with the MKF estimates of the population means or prevalence rates and their associated RMSE values). The variable names for the estimates and RMSEs depend on the models specified by the user. For the one-outcome default model, the variables are &outcome._pred_B and &outcome._se_B for the estimates and their RMSEs, where &outcome. is

the outcome variable name specified in the parameter outcome= and is used as a prefix for the estimate names. Variable names for other model specifications are given below.

system = optional operating system name

specifies the operating system being used for the analysis. The macro works in Windows© and Linux© operating systems. The options for this parameter are

- windows, the DEFAULT;
- linux.

by = optional variable name

by specifying the by parameter, the user can obtain separate analyses on observations in subsets defined by the by variables. Only one variable can be specified. When a value is given to the by parameter, the data structure needs to be identical for each value of the by parameter, i.e., each data subset. It must also include the by variable, which identifies the *BY* subsets. When a by parameter appears, the macro executes for the different subsets and outputs the data. A sorting of the data in order of the *BY* variable is not necessary. When the by parameter is left empty (the default), subset analyses are not conducted.

- DEFAULT = (empty).

Note we use (empty) to indicate that a macro parameter is specified and set to no value or by default is set to no value. For example, to set by to no value, the user might specify:

```
%mkf(
data= disease , group= group ,
time= time , outcome= y ,
se= se , software_dir= C: \SASAnalysis ,
by= , out= estimates
) ;
```

or

```
%mkf(
data= disease , group= group ,
time= time , outcome= y ,
se= se , software_dir= C: \SASAnalysis ,
by= %str() , out= estimates
) ;
```

When a user does not specify an optional macro parameter (as in the macro call in Section 2 in which by and all the other optional parameters are not specified), the parameter is set to its default as defined by the developers of the macro, which may or may not be no value (empty).

comparedto = optional group name

specifies the level of the group variable to be used as a comparison group when the differences between groups are of interest. The macro will produce estimates of the differences between the most recent year population mean for the comparison group and each other group. In the example in Section 2, because the group names in the data are {1, 2, 3, 4}, specifying comparedto = 3 will produce differences between the population mean of (group=3 and the mean of each of the other groups.

In Section 4.1, where race is the group studied, taking values {white, black, Chinese, Cuban, ...} specifying comparedto = black will produce differences between the population mean of Black and the mean of each of the other racial/ethnic groups. The comparedto parameter is not case-specific (e.g., black, Black or BLACK, or any variation will work). If the value of comparedto specified by the user does not match any of the values of the group variable in the data (e.g., if the user mistypes White as Whites), a WARNING message will be printed in the log file and no estimates of population mean differences will be computed. If the parameter is left empty (comparedto = (empty), the default), no difference estimation will be computed. See Section 4.1 for an example of the use of this option.

- DEFAULT = (empty).

Note: This option is available only when using the Bayesian method. It is not available when modeling two outcomes.

comparedata = optional SAS© data set name

specifies the name of the data where the difference estimates are stored. It is needed only if the comparedto parameter is specified.

- DEFAULT = &out._diff , a data name that uses the name specified in &out as a prefix and a suffix _diff.

slopes = optional keyword

specifies which, if any, MKF procedures using MLE should be performed. The valid values for this parameter are

- independent;

- common;
- dropped;
- or any combination of these three options;
- or (empty) (the default when using one outcome);
- or independent common (the default when using two outcomes).

The default for one outcome is to leave the parameter empty, which results in the use of the Bayesian approach alone. When the parameter is not empty, the keyword determines the assumptions about the linear time trends for the separate groups. The parameters of the given model specification are estimated via the maximum likelihood estimation procedure implemented in PROC NLMIXED. The model assumptions corresponding to the keywords are

1. independent assumes for each group a separate distinct slope parameter for the linear time trend;
2. common assumes a single slope parameter common to all groups;
3. dropped assumes no time trend for any group, i.e., the slope parameter is forced to zero for all groups.

When a combination of more than one option is specified (e.g., independent common), the macro produces MKF estimates for each of the requested methods and a model-averaged estimate is also produced. The model-averaging method produces a weighted average of the estimates from the component estimation method, where the weights are determined by the model fit of each component. See Section 5 for details on the model-averaging procedures. The RMSE estimates produced by the maximum likelihood estimation methods do not fully account for the estimation of all model parameters and might underestimate the error in the MKF predicted means. The Bayesian methods fully account for the estimation of all parameters and provide accurate RMSE estimates. We strongly encourage the use of the Bayesian procedures rather than the maximum likelihood procedures when analyzing only one outcome. Nonetheless, we discuss in Section 5 situations in which users might want to use the maximum likelihood-based approach. The variable names for the estimated population means and RMSE correspond to the estimation procedures specified by slopes according the following rules (where &outcome is the outcome name used as prefix).

Estimation variable label in output data			
Method Used	Options	Point Estimate	Standard Error
slopes	independent	&outcome_pred_G	&outcome_se_G
	common	&outcome_pred_1	&outcome_se_1
	dropped	&outcome_pred_0	&outcome_se_0
	Model Averaging	&outcome_pred_MA	&outcome_se_MA

bayesmodel = optional keyword

specifies which, if any, Bayesian estimation method is to be used to produce the MKF estimates of the population's means. This option is available only for one-outcome analysis. The options for this parameter are

- independent;
- common;
- full (the default);
- any combination of these 3 options;
- or (empty).

The model assumptions corresponding to the keywords are:

1. independent assumes a separate distinct slope for each group (the prior distribution for the slope's parameters assumes that they are independent, mean zero, with large non informative variances).
2. common assumes a single slope parameter common to all groups (the prior distribution for this slope parameter is mean zero, with large non informative variances).
3. full assumes separate slopes for each group but assumes that they are from a common distribution, which shrinks the estimates toward a common value (the prior distribution for the slope's parameters assumes that they are independent draws from a common normal distribution with mean zero and unknown variance, which is estimated from the data).

When a combination of more then one option is specified, all the specified options are estimated and when (empty) is specified, the Bayesian method will not be used. The variable names for the estimated population means and RMSE correspond to the estimation procedures specified by slopes according the following rules:

Estimation Variable Label in Output Data			
Method Used	Options	Point Estimate	Standard Error
bayesmodel	full	&outcome_pred_B	&outcome_se_B
	independent	&outcome_pred_BG	&outcome_se_BG
	common	&outcome_pred_B1	&outcome_se_B1

modelprint = keyword

specifies whether the PROC NLMIXED procedure estimates should be printed if the parameter slopes requests, MLE procedures be used. These model parameter estimates are manipulated through the Kalman filter that produce the desired population mean estimates (Kalman, 1960). These parameter values are not necessary for interpreting the final output but may be of interest to some users. The options for this parameter are

- no (the default);
- yes.

finalprint = keyword

specifies whether or not the final desired population estimates should be printed. If multiple options are specified in slopes and/or in bayesmodel, only the results of one estimation method will be printed in the SAS output window. All the other estimates will be saved in the data specified in the parameter out. When multiple estimation methods are used, as only one of the methods results will be printed, the priority order in which the chosen option estimates are to be printed in the SAS© output window is as follows:

1. bayesmodel = full;
2. bayesmodel = independent;
3. bayesmodel = common;
4. slopes = a combination of slopes options (model-averaging);
5. slopes = independent;
6. slopes = common;
7. slopes = dropped.

For example if both bayesmodel = full and slopes = independent common are specified, only the bayesmodel = full output will be printed in the SAS© window. The options for this parameter are

- no;
- yes (the default).

xtrakeep = optional variable name

the names of any variable in the user-supplied input data that the user wants to keep in the model estimate output data specified by out. The default is to keep the parameter value empty and include only the standard variables in the output.

pdigit = optional integer

an integer that specifies the number of decimal digits for the printed outputs. The default is set to 4.

4 Examples

This section provides two examples to demonstrate the default features and common optional features of the MKF macro.

4.1 Example 1: Prevalence of Stroke in the United States in Different Racial/Ethnic Groups, (1997-2004 NHIS Data)

The data cover the 1997-2004 $(T = 8)$ prevalence rates for stroke from the NHIS for $G =$ 11 different racial/ethnic groups (each year) in the United States (white, black, American Indian/ Alaskan Native, Chinese, Filipino, Asian Indian, Puerto Rican, Mexican, Cuban, other Hispanics and other racial/ethinic groups).

Our choice of the NHIS data set is based on its use as the primary sampling frame for many of the other National Center for Health Statistics (NCHS) data sets and the fact that NHIS contains the largest sample available of several small racial/ethnic groups over a long period of time. Since 1957, the NHIS has continuously conducted nationwide household interviews to collect information concerning the health of the U.S. civilian, non institutionalized population. The survey collects information on race/ethnicity, socioeconomic characteristics, and self-reported health status. There have been two basic redesign issues related to the NHIS over the years. The first one, an adjustment to the sampling design in 1995, led to state-level stratification, increasing the number of primary sampling locations from 198 to 358. This enhanced the capability of using the NHIS for state estimation and future dual-frame surveys at the state level. In addition, both the black and Hispanic populations are oversampled to allow for more precise estimation of health in these growing minority populations. The second one, in 1997, collected information on everyone in a sampled family, and the sample also served as a sampling frame for additional integrated surveys, a feature that was absent in earlier designs. The 1997-2004 adult sample contains similar information on race/ethnicity over time and also annually includes 200 complete cases each for such small groups as American Indian/Alaskan Native (AI/AN) and Chinese. See Elliott et al. (2009) for details on the definitions of the racial/ ethnic groups.

In the data set, the outcome of interest called STROKE is the proportion of stroke patients observed in the specified racial/ethnic group over the years. The rates and standard errors were calculated using the NHIS analysis weights and using standard procedures to account for the complex sampling design and nonresponse. The following SAS© statement creates the dataset. The variables *race* and *year* denote the racial/ethnic group and year of data.

```
data Prevalence;
length race $20;
label race =’Race group surveyed’
year=’Year of the survey’
stroke = ’Prevalence of Stroke’
se = ’Prevalence Standard Error’
;
input race $ year stroke se @@;
 datalines;
White     1997  0.02596  0.00106  Black       1997  0.02795  0.00232
White     1998  0.02589  0.00109  Black       1998  0.02794  0.00260
White     1999  0.02462  0.00110  Black       1999  0.03235  0.00277
White     2000  0.02489  0.00110  Black       2000  0.02840  0.00249
White     2001  0.02650  0.00111  Black       2001  0.03309  0.00273
White     2002  0.02735  0.00117  Black       2002  0.02544  0.00243
White     2003  0.02738  0.00120  Black       2003  0.03199  0.00281
White     2004  0.03199  0.00129  Black       2004  0.03012  0.00282
AI-AN     1997  0.04498  0.01500  Chinese     1997  0        0
AI-AN     1998  0.02335  0.01163  Chinese     1998  0.00604  0.00602
AI-AN     1999  0.04704  0.01630  Chinese     1999  0.00705  0.00703
AI-AN     2000  0.06062  0.01749  Chinese     2000  0.01622  0.00980
AI-AN     2001  0.01973  0.00919  Chinese     2001  0.00504  0.00503
AI-AN     2002  0.03463  0.01414  Chinese     2002  0.00672  0.00475
AI-AN     2003  0.03889  0.01466  Chinese     2003  0.01697  0.00972
AI-AN     2004  0.03767  0.01444  Chinese     2004  0.02096  0.01062
Filipino  1997  0        0        AsianIndian 1997  0.00473  0.00472
Filipino  1998  0.01988  0.01161  AsianIndian 1998  0.00695  0.00694
Filipino  1999  0.02983  0.01359  AsianIndian 1999  0        0
Filipino  2000  0.03474  0.01720  AsianIndian 2000  0        0
Filipino  2001  0.06470  0.02103  AsianIndian 2001  0        0
Filipino  2002  0.02951  0.01698  AsianIndian 2002  0.01219  0.00938
Filipino  2003  0.01893  0.01104  AsianIndian 2003  0        0
Filipino  2004  0.03093  0.01396  AsianIndian 2004  0.00697  0.00694
PuertoRican 1997  0.01101  0.00422  Mexican   1997  0.01058  0.00254
PuertoRican 1998  0.02821  0.00805  Mexican   1998  0.01134  0.00320
PuertoRican 1999  0.03095  0.00774  Mexican   1999  0.00826  0.00236
PuertoRican 2000  0.02324  0.00651  Mexican   2000  0.01260  0.00280
PuertoRican 2001  0.04042  0.00932  Mexican   2001  0.00867  0.00230
PuertoRican 2002  0.02876  0.00785  Mexican   2002  0.01169  0.00261
PuertoRican 2003  0.01770  0.00601  Mexican   2003  0.01030  0.00204
PuertoRican 2004  0.02454  0.00695  Mexican   2004  0.01228  0.00273
Cuban     1997  0.00899  0.00542  Other-Hisp  1997  0.01862  0.00281
Cuban     1998  0.00876  0.00443  Other-Hisp  1998  0.01608  0.00246
```

```
 Cuban      1999  0.00620  0.00362  Other-Hisp  1999  0.00892  0.00191
 Cuban      2000  0.01165  0.00961  Other-Hisp  2000  0.01248  0.00228
 Cuban      2001  0.03732  0.01354  Other-Hisp  2001  0.01679  0.00278
 Cuban      2002  0.01378  0.00798  Other-Hisp  2002  0.01675  0.00300
 Cuban      2003  0.02622  0.01158  Other-Hisp  2003  0.01139  0.00207
 Cuban      2004  0.03089  0.01535  Other-Hisp  2004  0.02224  0.00291
 All-Other  1997  0.00482  0.00296  All-Other   2001  0.02295  0.00702
 All-Other  1998  0.01383  0.00535  All-Other   2002  0.01190  0.00511
 All-Other  1999  0.00511  0.00361  All-Other   2003  0.01273  0.00536
 All-Other  2000  0.01169  0.00531  All-Other   2004  0.01928  0.00640
    ;
run;
```

As shown in these data, in 1997, 2.596% of whites reported having had a stroke with a standard error of 0.106%, and the prevalence rate increases over time to 3.199% in 2004. Among Chinese, no stroke cases were observed in the 1997 data, giving a prevalence of 0% with a standard error of 0%, although stroke cases were observed for this group from 1998 to 2004. Similarly, there were no stroke cases for Filipinos in 1997 or Asian Indians for 1999, 2000, 2001, and 2003. As noted above and described in detail below, the standard error of zero underestimates the true variability in the sample prevalence rates, and so the MKF macro will impute an alternative value using the nonzero values from the other years for each of these group years.

The MKF macro statement used to predict the prevalence rate for the most recent year (2004) is as follows:

```
%mkf(
        data= prevalence ,    group= race ,
        time= year ,          outcome= stroke ,
        se= se ,              software_dir= C:\SASAnalysis,
        out= results
        )  ;
```

This macro call specifies that the data set prevalence be used with the outcome variable stroke and the standard error se. The group and time parameters are respectively specified as race and year. The final required parameter software_dir establishes C:\SASAnalysis as the directory which contains the SAS© macro code and the statistical computation executable file. In addition, the call specifies that the output results should be collected in a data set called results. By default the MKF procedure will use full Bayesian model estimation procedures. The SAS© Windows© output is as follows:

MKF Full Bayesian Estimation for the outcome stroke

race	year	Estimation Type	Point Estimate	Std. Error	95% CI	Stdized Diff	Relative RMSE
White	2004	Sample	0.0320	0.0013	[0.0295, 0.0345]	--	--
		MKF estimate	0.0306	0.0012	[0.0282, 0.0330]	-1.0465	0.9463
Black	2004	Sample	0.0301	0.0028	[0.0246, 0.0356]	--	--
		MKF estimate	0.0309	0.0018	[0.0273, 0.0345]	0.2777	0.6515
AI-AN	2004	Sample	0.0377	0.0144	[0.0094, 0.0660]	--	--
		MKF estimate	0.0359	0.0053	[0.0255, 0.0463]	-0.1254	0.3675
Chinese	2004	Sample	0.0210	0.0106	[0.0001, 0.0418]	--	--
		MKF estimate	0.0106	0.0034	[0.0039, 0.0173]	-0.9735	0.3226
Warning: For this group, user supplied SE=0 were set to average of nonzero values							
Filipino	2004	Sample	0.0309	0.0140	[0.0036, 0.0583]	--	--
		MKF estimate	0.0275	0.0056	[0.0166, 0.0384]	-0.2460	0.3996
Warning: For this group, user supplied SE=0 were set to average of nonzero values							
AsianIndian	2004	Sample	0.0070	0.0069	[-0.0066, 0.0206]	--	--
		MKF estimate	0.0058	0.0033	[-0.0006, 0.0123]	-0.1640	0.4737
Warning: For this group, user supplied SE=0 were set to average of nonzero values							
PuertoRican	2004	Sample	0.0245	0.0070	[0.0109, 0.0382]	--	--
		MKF estimate	0.0248	0.0033	[0.0184, 0.0313]	0.0438	0.4717
Mexican	2004	Sample	0.0123	0.0027	[0.0069, 0.0176]	--	--
		MKF estimate	0.0122	0.0018	[0.0087, 0.0156]	-0.0415	0.6493
Cuban	2004	Sample	0.0309	0.0154	[0.0008, 0.0610]	--	--
		MKF estimate	0.0150	0.0040	[0.0072, 0.0229]	-1.0336	0.2614
Other-Hisp	2004	Sample	0.0222	0.0029	[0.0165, 0.0279]	--	--
		MKF estimate	0.0177	0.0020	[0.0137, 0.0217]	-1.5648	0.7025
All-Other	2004	Sample	0.0193	0.0064	[0.0067, 0.0318]	--	--
		MKF estimate	0.0142	0.0030	[0.0083, 0.0201]	-0.7960	0.4696

In the output, a warning message is given for the Chinese, Filipino, and Asian Indian groups to notify the users that this group had $SE = 0$ for one or more years and the macro replaced the zero values with imputed nonzero values using the standard errors from other years. In this example, the MKF procedure resulted in notable changes to the estimates of the prevalance of stroke in 2004 for several of the racial/ethnic groups. Inspection of the data reveals high rates in 2004 relative to earlier years for many of the groups, and the MKF method smooths out what is estimated to be sampling error, resulting in lower rates. For example, the MKF yields an estimate of the prevalence of stroke in 2004 for Chinese of 1%, whereas the sample rate was just over 2%. The goal of the MKF is to improve the accuracy of estimates for racial/ethnic groups with small NHIS sample sizes. The relative RMSE between MKF estimates and the sample estimates ranged from .2612 to .4737 for the American Indian/Alaskan Native, Chinese, Filipino, Asian Indian, Puerto Rican, and other groups, which all have small sample sizes in the NHIS. The MKF procedure produced an estimated standard error approximately 50% to 75% smaller than the observed standard error. These substantial gains are consistent with gains observed in simulations presented in Lockwood et al (2011), which could be of great value in applications of these estimates in health disparities and other policy research.

4.2 Disparities and Differences

Suppose the user wanted to study disparities in health outcomes by comparing all the other racial/ethnic groups to white (reference group) using the full Bayesian estimation method. The macro specification to produce these estimates is

```
%mkf(
                data= prevalence ,    group= race ,
                time= year ,          outcome= stroke ,
                 se= se ,             software_dir= C:\SASAnalysis,
                 out= results ,         comparedto= white
                )   ;
```

In this case the SAS© Windows© output will be as follows:

```
                        MKF Full Bayesian Estimation for the outcome
                                              stroke
        race         year  Estimation     Point       Std.   95% CI              Stdized  Relative
                             Type        Estimate    Error                       Diff     RMSE
#############################################################################################
        White        2004  Sample        0.0320      0.0013  [ 0.0295,  0.0345] --       --
                           MKF estimate  0.0306      0.0012  [ 0.0282,  0.0330] -1.0465  0.9463
---------------------------------------------------------------------------------------------
        Black        2004  Sample        0.0301      0.0028  [ 0.0246,  0.0356] --       --
                           MKF estimate  0.0309      0.0018  [ 0.0273,  0.0345] 0.2777   0.6515
---------------------------------------------------------------------------------------------
        AI-AN        2004  Sample        0.0377      0.0144  [ 0.0094,  0.0660] --       --
                           MKF estimate  0.0359      0.0053  [ 0.0255,  0.0463] -0.1254  0.3675
---------------------------------------------------------------------------------------------
        Chinese      2004  Sample        0.0210      0.0106  [ 0.0001,  0.0418] --       --
                           MKF estimate  0.0106      0.0034  [ 0.0039,  0.0173] -0.9735  0.3226
   Warning:  For this group, user supplied SE=0 were set to average of nonzero values
---------------------------------------------------------------------------------------------
        Filipino     2004  Sample        0.0309      0.0140  [ 0.0036,  0.0583] --       --
                           MKF estimate  0.0275      0.0056  [ 0.0166,  0.0384] -0.2460  0.3996
   Warning:  For this group, user supplied SE=0 were set to average of nonzero values
---------------------------------------------------------------------------------------------
        AsianIndian  2004  Sample        0.0070      0.0069  [-0.0066,  0.0206] --       --
                           MKF estimate  0.0058      0.0033  [-0.0006,  0.0123] -0.1640  0.4737
   Warning:  For this group, user supplied SE=0 were set to average of nonzero values
---------------------------------------------------------------------------------------------
        PuertoRican  2004  Sample        0.0245      0.0070  [ 0.0109,  0.0382] --       --
                           MKF estimate  0.0248      0.0033  [ 0.0184,  0.0313] 0.0438   0.4717
---------------------------------------------------------------------------------------------
        Mexican      2004  Sample        0.0123      0.0027  [ 0.0069,  0.0176] --       --
                           MKF estimate  0.0122      0.0018  [ 0.0087,  0.0156] -0.0415  0.6493
---------------------------------------------------------------------------------------------
        Cuban        2004  Sample        0.0309      0.0154  [ 0.0008,  0.0610] --       --
                           MKF estimate  0.0150      0.0040  [ 0.0072,  0.0229] -1.0336  0.2614
---------------------------------------------------------------------------------------------
        Other-Hisp   2004  Sample        0.0222      0.0029  [ 0.0165,  0.0279] --       --
                           MKF estimate  0.0177      0.0020  [ 0.0137,  0.0217] -1.5648  0.7025
---------------------------------------------------------------------------------------------
        All-Other    2004  Sample        0.0193      0.0064  [ 0.0067,  0.0318] --       --
                           MKF estimate  0.0142      0.0030  [ 0.0083,  0.0201] -0.7960  0.4696
---------------------------------------------------------------------------------------------
                          Differences between race MKF Point Estimates
         Compared Race                   Difference   Std.   95% CI
        AI-AN        - White             0.0052      0.0054  [-0.0054,  0.0158]
        All-Other    - White            -0.0165      0.0031  [-0.0226, -0.0103] **
        AsianIndian  - White            -0.0248      0.0035  [-0.0316, -0.0180] **
        Black        - White             0.0003      0.0022  [-0.0040,  0.0046]
        Chinese      - White            -0.0200      0.0035  [-0.0270, -0.0131] **
        Cuban        - White            -0.0156      0.0041  [-0.0236, -0.0076] **
        Filipino     - White            -0.0031      0.0057  [-0.0143,  0.0080]
        Mexican      - White            -0.0185      0.0021  [-0.0226, -0.0143] **
        Other-Hisp   - White            -0.0130      0.0021  [-0.0171, -0.0088] **
        PuertoRican  - White            -0.0058      0.0034  [-0.0125,  0.0009]
---------------------------------------------------------------------------------------------
---------------------------------------------------------------------------------------------
```

A difference estimate between all the groups and white is printed at the end of the output. The standard errors as well as 95% confidence intervals are included.

If the user specified the parameter comparedto= white as comparedto= whites with an "s" at the end of white, then it will not match any values of the group variable. No difference estimates will be generated and the following warning message will be shown in the SAS© log file:

```
Warning: The comparison group Whites is not a race value.
Warning: Check to make sure the value Whites is correct.
Warning: No comparison will be printed at this point. All comparisons
  could be found in the param_bayes data
```

4.3 Nonconform Data Error Messages

As discussed above, the MKF macro does not allow for missing values in the outcome or standard errors. In the following example, we set the value of SE to missing for whites in 2000 and rerun the macro:

```
data prevalence2;
 set prevalence;
 if race="White" and year=2000 then se=.;
run;

%MKF(
      data=prevalence2,      group=race,
      time=year,             outcome=stroke,
      se=se,                 Software_Dir= C:\SASAnalysis,
      Out=Results2
      )
```

And the output produces the following error message:

```
                                    Error Note:

                              An error occurred with your data.

       Check the data and make sure that there is no missing value in the outcome and SE and

            that all the groups have exactly the same number of year(time) worth of data
```

The user will receive this error message notification whenever the input data set is not properly entered in the required format. In this event, the user should check the data to identify the errors and correct them before rerunning the MKF macro.

4.4 The Output Data File

In addition to creating the printout table of results, the macro also creates an output data file that contains the estimated population means or prevalence rates and their associated RMSE for each group and each MKF model option specified in the macro. In the following example, the MLE options of independent and common slope are specified (via the slopes parameter) in addition to the full Bayesian model fit, which must be requested through the specification of the bayesmodel parameter because the slopes parameter is not empty:

```
%mkf(
        data= prevalence ,     group= race ,
        time= year ,           outcome= stroke ,
        se= se ,               software_dir= C: \SASAnalysis,
        slopes= independent common ,      bayesmodel= full ,
        out= results
        )   ;
```

Because the macro gives preference to the Bayesian full method, the SAS© output window provides only the full Bayesian estimates, but all the other estimates can be recovered from the data output with the parameter Out=Results . Here is what the first 15 lines of the output look like:

```
race          year     stroke       se       pred_MA       rmse_MA     pred_G
AI-AN         1997    0.04498    0.01500    0.031709    .009038903    0.034528
AI-AN         1998    0.02335    0.01163    0.032027    .007299822    0.034095
AI-AN         1999    0.04704    0.01630    0.032938    .005870732    0.034116
AI-AN         2000    0.06062    0.01749    0.033356    .005022232    0.033762
AI-AN         2001    0.01973    0.00919    0.033621    .004935953    0.033279
AI-AN         2002    0.03463    0.01414    0.034640    .005863410    0.033377
AI-AN         2003    0.03889    0.01466    0.034963    .007277750    0.032955
AI-AN         2004    0.03767    0.01444    0.035672    .008985126    0.032808
All-Other     1997    0.00482    0.00296    0.007972    .002449589    0.005597
All-Other     1998    0.01383    0.00535    0.009679    .002525411    0.007668
All-Other     1999    0.00511    0.00361    0.008892    .001933777    0.008506
All-Other     2000    0.01169    0.00531    0.010789    .002102774    0.010971
All-Other     2001    0.02295    0.00702    0.010867    .002301006    0.012361
All-Other     2002    0.01190    0.00511    0.011398    .002638974    0.013818
All-Other     2003    0.01273    0.00536    0.012002    .003206071    0.015550

     rmse_G     pred_1         rmse_1     pred_B         rmse_B
.009021375    0.031709    .009038903    0.031686    .005298438
.007276538    0.032027    .007299822    0.031999    .005108730
.005841759    0.032938    .005870732    0.032874    .005033377
.004979832    0.033356    .005022232    0.033569    .004996362
.004910078    0.033621    .004935953    0.033548    .004891341
.005828613    0.034640    .005863410    0.034587    .005032366
.007252550    0.034963    .007277750    0.035210    .005136055
.008969690    0.035672    .008985126    0.035859    .005306086
.002436454    0.007972    .002449589    0.007071    .002188699
.002407460    0.009679    .002525411    0.009035    .002420796
.001883343    0.008892    .001933777    0.008792    .002082060
.002015942    0.010789    .002102774    0.010537    .002256394
.002228046    0.010867    .002301006    0.011789    .002432417
.002588552    0.011398    .002638974    0.012164    .002445783
.003150004    0.012002    .003206071    0.013010    .002670158
......
......
```

The estimates are provided for each group and every year. The MLE procedures specified by the slope parameters produce estimates and RMSE for a model with separate slopes for every group (pred_G and rmse_G), a model with a common slope for all groups (pred_1 and rmse_1), and a model average estimate that averages the two other estimates (pred_MA and rmse_MA). The results for the full Bayesian procedure are given in pred_B and rmse_B.

4.5 Example 2: Subset Analysis of the Prevalence of a Disease Among Men and Women in Different Racial/Ethnic Groups

In this example with simulated data, the goal is to estimate the prevalence of a disease in difference racial/ethnic groups by gender. Here is the data set:

```
 data withgender;
 length Race $20 Gender$20;
 label Gender ='Gender subset'
 Race ='Race group surveyed'
 Year='Year of the survey'
 Disease = 'Prevalence of the Disease'
 SE = 'Prevalence Standard Error'
 ;
 input Gender $ Race $ Year Disease SE @@;
  datalines;
 Male   White   2000  0.2455 0.0028  Male   White   2001  0.2454 0.0030
 Male   White   2002  0.2461 0.0031  Male   White   2003  0.2464 0.0031
 Female White   2000  0.1197 0.0231  Female White   2001  0.1239 0.0273
 Female White   2002  0.1109 0.0252  Female White   2003  0.1240 0.0259
 Male   Black   2000  0.2527 0.0031  Male   Black   2001  0.2650 0.0032
 Male   Black   2002  0.2744 0.0033  Male   Black   2003  0.2794 0.0033
 Female Black   2000  0.1309 0.0255  Female Black   2001  0.1914 0.0293
 Female Black   2002  0.1475 0.0270  Female Black   2003  0.1712 0.0309
 Male   Chinese 2000  0.3138 0.0068  Male   Chinese 2001  0.3221 0.0075
 Male   Chinese 2002  0.3063 0.0074  Male   Chinese 2003  0.3141 0.0072
 Female Chinese 2000  0.1666 0.0076  Female Chinese 2001  0.1804 0.0079
 Female Chinese 2002  0.1810 0.0085  Female Chinese 2003  0.1634 0.0082
 Male   Indian  2000  0.3232 0.0073  Male   Indian  2001  0.3215 0.0076
 Male   Indian  2002  0.3439 0.0079  Male   Indian  2003  0.3334 0.0077
 Female Indian  2000  0.1249 0.0086  Female Indian  2001  0.1227 0.0096
 Female Indian  2002  0.1161 0.0084  Female Indian  2003  0.1300 0.0086
 Male   Hispanic 2000 0.2530 0.0373  Male   Hispanic 2001 0.2263 0.0388
 Male   Hispanic 2002 0.2100 0.0355  Male   Hispanic 2003 0.2852 0.0383
 Female Hispanic 2000 0.1173 0.0289  Female Hispanic 2001 0.1083 0.0276
 Female Hispanic 2002 0.1055 0.0245  Female Hispanic 2003 0.1056 0.0226
 Male   Other   2000  0.2696 0.0314  Male   Other   2001  0.2686 0.0353
 Male   Other   2002  0.2597 0.0354  Male   Other   2003  0.3326 0.0364
 Female Other   2000  0.0649 0.0194  Female Other   2001  0.1304 0.0269
 Female Other   2002  0.1993 0.0178  Female Other   2003  0.2232 0.0184
    ;
run;
```

The data include sample mean estimates and standard errors by gender and racial/ethnic group for every year. They also include the variable gender. Separate estimation by gender is specified in the macro through the value of the by parameter (by = gender). The macro call is

```
%mkf(
        data= withgender ,    group= race ,
        time= year ,          outcome= disease ,
        se= se ,              software_dir= C:\SASAnalysis,
        by= gender ,                out= results
        ) ;
```

The output from this analysis prints the the result for men and women separately:

MKF Full Bayesian Estimation for the outcome Disease

gender	race	year	Estimation Type	Point Estimate	Std. Error	95% CI	Stdized Diff	Relative RMSE
Male	White	2003	Sample	0.2464	0.0031	[0.2403, 0.2525]	--	--
			MKF estimate	0.2468	0.0027	[0.2414, 0.2522]	0.1226	0.8839
Female	White	2003	Sample	0.1240	0.0259	[0.0732, 0.1748]	--	--
			MKF estimate	0.1224	0.0206	[0.0820, 0.1628]	-0.0623	0.7963
Male	Black	2003	Sample	0.2794	0.0033	[0.2729, 0.2859]	--	--
			MKF estimate	0.2801	0.0030	[0.2743, 0.2859]	0.2069	0.9018
Female	Black	2003	Sample	0.1712	0.0309	[0.1106, 0.2318]	--	--
			MKF estimate	0.1717	0.0233	[0.1261, 0.2174]	0.0176	0.7543
Male	Chinese	2003	Sample	0.3141	0.0072	[0.3000, 0.3282]	--	--
			MKF estimate	0.3144	0.0060	[0.3027, 0.3260]	0.0365	0.8278
Female	Chinese	2003	Sample	0.1634	0.0082	[0.1473, 0.1795]	--	--
			MKF estimate	0.1693	0.0076	[0.1545, 0.1841]	0.7207	0.9212
Male	Indian	2003	Sample	0.3334	0.0077	[0.3183, 0.3485]	--	--
			MKF estimate	0.3372	0.0061	[0.3252, 0.3491]	0.4886	0.7937
Female	Indian	2003	Sample	0.1300	0.0086	[0.1131, 0.1469]	--	--
			MKF estimate	0.1267	0.0077	[0.1117, 0.1418]	-0.3819	0.8919
Male	Hispanic	2003	Sample	0.2852	0.0383	[0.2101, 0.3603]	--	--
			MKF estimate	0.2494	0.0216	[0.2070, 0.2918]	-0.9349	0.5644
Female	Hispanic	2003	Sample	0.1056	0.0226	[0.0613, 0.1499]	--	--
			MKF estimate	0.1074	0.0189	[0.0704, 0.1443]	0.0774	0.8348
Male	Other	2003	Sample	0.3326	0.0364	[0.2613, 0.4039]	--	--
			MKF estimate	0.2915	0.0210	[0.2504, 0.3326]	-1.1301	0.5760
Female	Other	2003	Sample	0.2232	0.0184	[0.1871, 0.2593]	--	--
			MKF estimate	0.2268	0.0163	[0.1948, 0.2588]	0.1957	0.8865

4.6 Example 3: Two outcomes, Prevalence of Hypertension and Diabetes in the United States in Different Racial/Ethnic Groups (1997-2004 NHIS Data).

This third example is also from the NHIS data, where for $G = 11$ different racial/ethnic groups in the United States (white, black, American Indian/Alaskan Native, Chinese, Filipino, Asian Indian, Puerto Rican, Mexican, Cuban, other Hispanics and other racial/ethnic groups), hypertension and diabetes are analyzed. It is well know that there is a relation between hypertension and diabetes, and this analysis will pool information from groups and time as well as across outcomes to estimate the prevalence of the two diseases in the different racial/ethnic groups. In the data set, the variable hyperten identifies the proportion of hypertension patients observed in the specified racial/ethnic group over the years, and the variable diabetes does the same for the proportion of diabetes patients. The standard errors of the sample prevalence rates are called hyperten_se and diabetes_se, respectively. The rates and standard errors were again calculated using the NHIS analysis weights and following standard procedures to account for the complex sampling design and nonresponse. The following SAS© statement creates the data. The variables *RACE* and *YEAR* denote the racial/ethnic group and year of data.

```
data Prevalences;
length race $20;
label race ='Race group surveyed'
year='Year of the survey'
Hyperten = 'Prevalence of Hypertension'
hyperten_se = 'Hypertension Prevalence Standard Error'
diabetes = 'Prevalence of Diabetes'
diabetes_se = 'Hypertension Diabetes Standard Error'
;
input race $ year stroke se @@;
 datalines;
Race Year Hyperten hyperten_se diabetes diabetes_se
White            1997    0.24552    0.00285    0.04916    0.00143
White            1998    0.24545    0.00301    0.04995    0.00151
White            1999    0.24612    0.00310    0.05226    0.00159
White            2000    0.24636    0.00307    0.05614    0.00164
White            2001    0.25272    0.00306    0.06034    0.00167
White            2002    0.26500    0.00320    0.06424    0.00176
White            2003    0.27438    0.00329    0.06599    0.00182
White            2004    0.27944    0.00329    0.06839    0.00183
Black            1997    0.31383    0.00681    0.08525    0.00408
Black            1998    0.32212    0.00752    0.08162    0.00422
Black            1999    0.30631    0.00745    0.08443    0.00445
Black            2000    0.31406    0.00717    0.09132    0.00439
Black            2001    0.32324    0.00733    0.09945    0.00467
Black            2002    0.32150    0.00755    0.09043    0.00455
Black            2003    0.34386    0.00786    0.09539    0.00470
Black            2004    0.33340    0.00765    0.10341    0.00493
AI/AN            1997    0.25404    0.03283    0.13213    0.02522
AI/AN            1998    0.22762    0.03404    0.07967    0.02030
AI/AN            1999    0.27299    0.03458    0.10109    0.02305
AI/AN            2000    0.28872    0.03318    0.11713    0.02334
AI/AN            2001    0.26955    0.03145    0.10618    0.02176
AI/AN            2002    0.26859    0.03531    0.11678    0.02484
AI/AN            2003    0.25972    0.03540    0.13320    0.02741
AI/AN            2004    0.33258    0.03638    0.16044    0.02758
Chinese          1997    0.11973    0.02309    0.01479    0.00900
Chinese          1998    0.12395    0.02729    0.03767    0.01585
Chinese          1999    0.11088    0.02519    0.03373    0.01568
Chinese          2000    0.12403    0.02591    0.03598    0.01492
Chinese          2001    0.13466    0.02625    0.04288    0.01664
Chinese          2002    0.13090    0.02546    0.05535    0.02096
Chinese          2003    0.19143    0.02932    0.06497    0.01870
Chinese          2004    0.14752    0.02700    0.06336    0.01911
Filipino         1997    0.17118    0.03091    0.04002    0.01634
Filipino         1998    0.25295    0.03727    0.06613    0.02087
```

Filipino	1999	0.22628	0.03882	0.04276	0.01531
Filipino	2000	0.21005	0.03553	0.03478	0.01437
Filipino	2001	0.28519	0.03825	0.04924	0.01575
Filipino	2002	0.18942	0.03639	0.06352	0.01936
Filipino	2003	0.23486	0.03382	0.05894	0.01870
Filipino	2004	0.24853	0.03280	0.08275	0.01978
Asian Indian	1997	0.09290	0.02440	0.04535	0.01716
Asian Indian	1998	0.05256	0.01958	0.02007	0.01157
Asian Indian	1999	0.11726	0.02895	0.08368	0.02641
Asian Indian	2000	0.10826	0.02762	0.04534	0.02030
Asian Indian	2001	0.10554	0.02447	0.03750	0.01425
Asian Indian	2002	0.10559	0.02262	0.03919	0.01535
Asian Indian	2003	0.06488	0.01938	0.05225	0.01839
Asian Indian	2004	0.13040	0.02691	0.10471	0.02425
Puerto Rican	1997	0.19929	0.01778	0.09176	0.01489
Puerto Rican	1998	0.22325	0.01841	0.10409	0.01343
Puerto Rican	1999	0.22352	0.01916	0.07403	0.01127
Puerto Rican	2000	0.21452	0.01914	0.09368	0.01325
Puerto Rican	2001	0.27312	0.02057	0.10737	0.01429
Puerto Rican	2002	0.21591	0.01970	0.08843	0.01291
Puerto Rican	2003	0.27689	0.02094	0.10752	0.01408
Puerto Rican	2004	0.26164	0.02143	0.10383	0.01448
Mexican	1997	0.12493	0.00860	0.04490	0.00517
Mexican	1998	0.12273	0.00962	0.04824	0.00641
Mexican	1999	0.11605	0.00844	0.05660	0.00618
Mexican	2000	0.13003	0.00862	0.05456	0.00547
Mexican	2001	0.13674	0.00829	0.05422	0.00536
Mexican	2002	0.14559	0.00899	0.06617	0.00668
Mexican	2003	0.13153	0.00775	0.04830	0.00490
Mexican	2004	0.15839	0.00881	0.06199	0.00548
Cuban	1997	0.24553	0.02669	0.06079	0.01351
Cuban	1998	0.17462	0.02363	0.05596	0.01354
Cuban	1999	0.28682	0.02858	0.06307	0.01477
Cuban	2000	0.25519	0.02638	0.06734	0.01538
Cuban	2001	0.29372	0.02922	0.05992	0.01416
Cuban	2002	0.25386	0.02806	0.10846	0.02127
Cuban	2003	0.25239	0.02707	0.05898	0.01357
Cuban	2004	0.29817	0.02965	0.12184	0.02242
Other Hispanic	1997	0.16660	0.00763	0.06034	0.00531
Other Hispanic	1998	0.18038	0.00791	0.06452	0.00481
Other Hispanic	1999	0.18101	0.00845	0.06913	0.00581
Other Hispanic	2000	0.16339	0.00821	0.06976	0.00574
Other Hispanic	2001	0.19683	0.00852	0.07484	0.00540
Other Hispanic	2002	0.19231	0.00858	0.05952	0.00488
Other Hispanic	2003	0.18214	0.00852	0.06801	0.00538
Other Hispanic	2004	0.17616	0.00787	0.07589	0.00542
All Other	1997	0.14503	0.01639	0.03528	0.00830
All Other	1998	0.16962	0.01799	0.04456	0.00974
All Other	1999	0.14074	0.01866	0.03200	0.00905
All Other	2000	0.18134	0.02039	0.03262	0.00882
All Other	2001	0.16428	0.01762	0.04135	0.00932
All Other	2002	0.14498	0.01662	0.06039	0.01187
All Other	2003	0.15313	0.01728	0.04217	0.00931
All Other	2004	0.19026	0.01868	0.05188	0.01086

So, in 2000, 24.64% of whites reported having hypertension with a standard error of 0.31%, and at the same time, the rate of diabetes was 5.61% with a standard error of 0.16%. The MKF macro statement used to predict the prevalence rate for both hypertension and diabetes for the most recent year (2004) is as follows:

```
%mkf(
        data= prevalences ,    group= race ,     time= year,
     outcome= hypertens ,     se= hypertens_se ,
     outcome2= diabetes ,     se2= diabetes_se ,
      software_dir= C: \SASAnalysis,          out= results
          )   ;
```

In this case, the SAS© Windows© output will be as follows:

MKF race slope MLE Model Averaging Estimation for the outcome
hyperten and diabetes

race	year	Estimation Type	Point Estimate	Std. Error	95% CI	Stdized Diff	Relative RMSE
White	2004	hyperten Estimations:					
		Sample	0.2794	0.0086	[0.2626, 0.2963]	--	--
		MKF estimate	0.2735	0.0061	[0.2615, 0.2854]	-0.6929	0.7107
		diabetes Estimations:					
		Sample	0.0684	0.0013	[0.0658, 0.0710]	--	--
		MKF estimate	0.0681	0.0010	[0.0663, 0.0700]	-0.1906	0.7107
Black	2004	hyperten Estimations:					
		Sample	0.3334	0.0229	[0.2885, 0.3783]	--	--
		MKF estimate	0.3375	0.0145	[0.3090, 0.3659]	0.1776	0.6333
		diabetes Estimations:					
		Sample	0.1034	0.0036	[0.0964, 0.1104]	--	--
		MKF estimate	0.1010	0.0023	[0.0966, 0.1055]	-0.6613	0.6333
AI/AN	2004	hyperten Estimations:					
		Sample	0.3326	0.1271	[0.0835, 0.5816]	--	--
		MKF estimate	0.2865	0.0779	[0.1337, 0.4392]	-0.3627	0.6133
		diabetes Estimations:					
		Sample	0.1604	0.0199	[0.1214, 0.1995]	--	--
		MKF estimate	0.1245	0.0122	[0.1006, 0.1484]	-1.8042	0.6133
Chinese	2004	hyperten Estimations:					
		Sample	0.1475	0.0883	[-0.0256, 0.3206]	--	--
		MKF estimate	0.1492	0.0536	[0.0442, 0.2542]	0.0192	0.6066
		diabetes Estimations:					
		Sample	0.0634	0.0138	[0.0362, 0.0905]	--	--
		MKF estimate	0.0476	0.0084	[0.0312, 0.0641]	-1.1367	0.6066
Filipino	2004	hyperten Estimations:					
		Sample	0.2485	0.0922	[0.0678, 0.4293]	--	--
		MKF estimate	0.2398	0.0568	[0.1285, 0.3510]	-0.0951	0.6156
		diabetes Estimations:					
		Sample	0.0828	0.0145	[0.0544, 0.1111]	--	--
		MKF estimate	0.0616	0.0089	[0.0441, 0.0790]	-1.4649	0.6156
Asian Indian	2004	hyperten Estimations:					
		Sample	0.1304	0.1111	[-0.0873, 0.3481]	--	--
		MKF estimate	0.1060	0.0559	[-0.0036, 0.2155]	-0.2199	0.5033
		diabetes Estimations:					
		Sample	0.1047	0.0174	[0.0706, 0.1388]	--	--
		MKF estimate	0.0535	0.0088	[0.0363, 0.0706]	-2.9439	0.5033
Puerto Rican	2004	hyperten Estimations:					
		Sample	0.2616	0.0671	[0.1302, 0.3931]	--	--
		MKF estimate	0.2493	0.0422	[0.1666, 0.3321]	-0.1836	0.6293
		diabetes Estimations:					
		Sample	0.1038	0.0105	[0.0832, 0.1244]	--	--
		MKF estimate	0.1046	0.0066	[0.0917, 0.1176]	0.0757	0.6293
Mexican	2004	hyperten Estimations:					
		Sample	0.1584	0.0255	[0.1084, 0.2084]	--	--
		MKF estimate	0.1476	0.0165	[0.1153, 0.1800]	-0.4216	0.6480
		diabetes Estimations:					
		Sample	0.0620	0.0040	[0.0542, 0.0698]	--	--
		MKF estimate	0.0630	0.0026	[0.0579, 0.0681]	0.2558	0.6480
Cuban	2004	hyperten Estimations:					
		Sample	0.2982	0.1033	[0.0957, 0.5007]	--	--
		MKF estimate	0.2685	0.0516	[0.1673, 0.3697]	-0.2871	0.4998
		diabetes Estimations:					
		Sample	0.1218	0.0162	[0.0901, 0.1536]	--	--
		MKF estimate	0.0782	0.0081	[0.0624, 0.0941]	-2.6931	0.4998
Other Hispanic	2004	hyperten Estimations:					
		Sample	0.1762	0.0251	[0.1270, 0.2253]	--	--
		MKF estimate	0.1939	0.0162	[0.1621, 0.2256]	0.7062	0.6461
		diabetes Estimations:					
		Sample	0.0759	0.0039	[0.0682, 0.0836]	--	--
		MKF estimate	0.0769	0.0025	[0.0719, 0.0819]	0.2543	0.6461
All Other	2004	hyperten Estimations:					
		Sample	0.1903	0.0508	[0.0908, 0.2897]	--	--
		MKF estimate	0.1750	0.0308	[0.1146, 0.2353]	-0.3014	0.6066
		diabetes Estimations:					
		Sample	0.0519	0.0080	[0.0363, 0.0675]	--	--
		MKF estimate	0.0513	0.0048	[0.0418, 0.0607]	-0.0750	0.6066

5 Details and Theory

This section provides details on the statistical models and estimation methods used by the MKF procedures and implemented in the MKF macro. In particular, it provides details on the Bayesian specification, the imputations of standard errors when the observed values are zero, and model-averaging for the likelihood estimation proecedures.

5.1 Data Model

The model is for data consisting of mean outcome measurements, Y_{gt} and X_{gt} (e.g., health status such as prevalence of diabetes, hypertension, or mean BMI), from $g = 1,\ldots,G$ racial/ethnic groups for $t = 1,\ldots,T$ time points. The data also include the standard errors of the means. The means could be from simple random samples or complex multistage samples of the racial/ethnic groups, and the standard errors of the means are assumed to be unbiased and to have accounted for the sample design in their calculation. The model for the analysis of one outcome Y_{gt} is

$$Y_{gt} = \alpha_g + \beta_g t + \gamma_{gt} + \varepsilon_{gt} \quad \text{with} \quad \varepsilon_{gt} \sim N(0, \sigma^2_{gt}). \tag{1}$$

When two outcomes are used in a multivariate fashion, the model used is:

$$\begin{aligned} Y_{gt} &= \alpha_g + \beta_g t + \gamma_{gt} + \varepsilon_{gt} \quad \text{with} \quad \varepsilon_{gt} \sim N(0, \sigma^2_{gt}) \\ X_{gt} &= \kappa_g + \lambda_g t + \phi_{gt} + \zeta_{gt} \quad \text{with} \quad \zeta_{gt} \sim N(0, \nu^2_{gt}). \end{aligned}$$

The random variables γ_{gt} and ϕ_{gt} are assumed to be autoregressive processes of order 1 (AR(1)) with the same autocorrelation coefficient ρ

$$\begin{aligned} \gamma_{gt} &= \rho\,\gamma_{g,t-1} + \xi_{gt} \quad \text{with} \quad \xi_{gt} \sim N(0, \tau^2) \\ \phi_{gt} &= \rho\,\phi_{g,t-1} + \nu_{gt} \quad \text{with} \quad \nu_{gt} \sim N(0, \delta\tau^2). \end{aligned}$$

For the outcome Y_{gt}, the true states of the mean outcomes for group g are given by a group-specific linear trend $(\alpha_g + \beta_g t)$ and a deviation from the trend at time t given by γ_{gt}. The unobserved true state of the group g mean outcome at time t is thus $\eta_{gt} = \alpha_g + \beta_g t + \gamma_{gt}$. The observed means deviate from the true states due to sampling error ε_{gt}, which we assume is normally distributed with mean zero and known variance σ^2_{gt} depending on the survey design and the effective sample size for group g at time t; we implicitly treat σ^2_{gt} as if it were known throughout. We assume that the ε_{gt} are independent both across groups and within groups across time. Thus we have

$$Y_{gt} \mid \alpha_g, \beta_g, \gamma_{gt} \sim \text{ind } N(\eta_{gt}, \sigma^2_{gt}), \tag{2}$$

and the likelihood function for the data conditional on the parameters (including the unobserved group-specific trends and the processes of deviations from those trend lines) is given by the product of Equation 2 across both groups and time periods.

When pooling information from a second outcome X_{gt} that is hypothesized to have the same structure with a group-specific linear trend $\kappa_g + \lambda_g t$ and a deviation from such trend at time t, we will assume that the deviation in the outcome X_{gt} is ϕ_{gt} with the same autocorrelation parameter ρ but similar innovation parameter τ^2 up to a multiplicative constant δ. Information from the outcome X_{gt} is then "borrowed" for the estimation of the correlation ρ. A full parameterization allowing for separate estimates of each AR(1) process for the different outcomes separately would be a more completed model specification where a correlation between the two can be estimated. But with the MKF dealing with only a small number of observations, estimation of these parameters separately would have been inefficient. In the proposed setup, the unique AR(1) correlation coefficients and different innovation parameters will be estimated and the Kalman filter algorithm will be used to estimate γ_{gt} from Y_{gt} and ϕ_{gt} directly from X_{gt}.

5.2 Prior Distributions for the Bayesian Models

In the Bayesian framework, the likelihood function is combined with prior distributions for the unknown parameters to create the posterior distribution for the unknown parameters given the data, from which all inferences are derived. In practical cases, features of the posterior distribution are estimated by sampling from the posterior distribution using Markov Chain Monte Carlo (MCMC) methods, (Carlin and Louis, 2000; Gelman et al., 1995; and Gilk et al., 1996). For this model, the unknown parameters consist of the "stochastic parents" (Lunn et al., 2000) of the observed data, i.e., α_g, β_g, γ_{gt}, as well as additional unknown parameters introduced below, which are stochastic parents of these parameters. The main parameters of interest are the unknown states η_{gt}, and in particular the unknown states η_{gT}, from the most recently observed (current) time point. These unobserved states are a deterministic function of the unknown parameters α_g, β_g, and γ_{gt}, and thus their posterior distribution is derived from the joint posterior distribution of all unknown parameters.

The prior distributions for the model are specified as follows. Unless otherwise indicated, components are assumed to be independent. Parameters of the prior distributions that are chosen fixed constants (i.e., not stochastic) are notated $c_1, c_2, \ldots$ Specific values of these constants are discussed in Section 5.4.

5.2.1 Group Intercepts and Slopes

The group intercepts α_g are modeled as independent $N(c_1, c_2)$.

In our primary model (the model fit when the macro parameter bayesmodel = full), the group slopes β_g are modeled as independent $N(\mu_\beta, v_\beta)$ conditional on their stochastic parent parameters, the mean slope for all groups μ_β, and the group-to-group variance in slopes v_β. These are unknown parameters informed by the data. Because the parent parameters are modeled as unknowns, prior distributions for them must be specified.

The mean slope μ_β is modeled as $N(c_3, c_4)$. The slope standard deviation across groups $\sqrt{v_\beta}$ is modeled as $U(c_5, c_6)$, where $U(a,b)$ is a uniform distribution with lower bound a and upper bound b.

The macro also allows users to select two alternative models by setting bayesmodel to common or dropped. These alternative models change the assumptions about the slope parameters but otherwise the model including the prior specifications remains the same. When bayesmodel = common, the group slopes β_g are assumed equal to a common unknown slope $\beta \sim N(c_7, c_8)$. When bayesmodel = dropped, $\beta_g = 0$ for all groups.

5.2.2 True State Deviations from Linear Trends

The state space deviations γ_{gt} are modeled with an AR(1) process within groups and are assumed to be independent across groups:

$$\begin{aligned} \gamma_{gt} &= \rho\gamma_{g,t-1} + \xi_{gt} \\ \xi_{gt} &\sim \text{iid } N(0, \tau^2). \end{aligned} \tag{3}$$

We assume the stationary version of this process so that the vector of deviations $\boldsymbol{\gamma}_g$ within a group has mean vector $\mathbf{0}$ and covariance matrix:

$$\text{Var}(\boldsymbol{\gamma}_g) = \frac{\tau^2}{1-\rho^2}\boldsymbol{A}(\rho), \tag{4}$$

where $\boldsymbol{A}(\rho)$ is the autoregressive correlation matrix with (i,j) entry $\rho^{|i-j|}$. The autocorrelation parameter ρ and innovation variance τ^2 are unknown parameters assumed to be common across groups.

The prior distribution for the autocorrelation parameter ρ is modeled through the transformation $\psi = \log(\frac{1-\rho}{1+\rho})$ with inverse transformation $\rho = \frac{1-e^\psi}{1+e^\psi}$. The parameter ψ is thus unbounded and maps to ρ in $(-1,1)$. The prior distribution for ψ is $N(c_9, c_{10})$. For the square root τ of the innovation variance, we used the prior distribution $U(c_{11}, c_{12})$.

5.3 Implementation

We developed an MCMC algorithm to calculate the posterior distributions of the model parameters including the states of the group means for the current year. We implemented the algorithm in the C/C++ programming language and have included it in the executable program file, which is controlled via the SAS© macro interface. Our MCMC algorithm used Gibbs sampling steps for α_g, β_g, $\boldsymbol{\gamma}_g$ and μ_β and Metropolis-Hastings accept/reject sampling steps for v_β, ρ, and τ^2 [4].

[4] We verified that our code was correct by checking that it produced identical inferences to those produced in Bayesian modeling software WinBUGS (Lunn et al., 2000) for a small number of the simulated

5.4 Specification of Prior Distributions

Our MKF software is designed to work with any possible nonnegative outcome, regardless of its scale (e.g., proportions on a (0,1) scale for dichotomous individual-level outcomes as well as continuous outcomes such as BMI). To allow for minimal user inputs in the specification of the Bayesian prior distributions, we defined our prior distributions as a function of the range r of the observed data (i.e., the maximum minus the minimum observed outcome for all groups and time points), which allows the prior distributions to be sensible regardless of the scale of the data. Although there are general concerns about potential bias in inferences when the data are used to inform the prior disrtibution, the very coarse use of the information in the data to calibrate the scale of the prior distributions had negligible effects on the substantive results of an extensive simulation study we conducted of the properties of our estimation method.

Table 2: List of fixed constants of prior distributions used in MKF software implementation of the Bayesian model (bayesmodel =full).

c_1	$0.5r$
c_2	$1000000r^2$
c_3	0.0
c_4	$0.1r^2$
c_5	0.0
c_6	$0.5r$
c_9	0.0
c_{10}	1.0
c_{11}	0.0001
c_{12}	$0.1r$

Table 2 provides the specification of the fixed constants for the prior distributions used in the MKF software. The prior mean c_1 of the α_g is set to $0.5r$, a rough approximation to the median of the data. However the large prior variance $c_2 = 1000000r^2$ essentially means that the intercepts are determined by the data. The mean slope c_3 is set to zero, reflecting no prior orientation to positive or negative slopes. The prior variance $c_4 = 0.1r^2$ for the mean slope indicates that large average growth relative to the scale of the data is not likely. The prior distribution for the standard deviation of the slopes across groups is $U(c_5 = 0, c_6 = 0.5r)$. One concern with using this type of prior is that if the upper bound is too low, the prior might have undue influence on the estimate of the heterogeneity of the slopes; in particular, it might result in underestimation of the variability

data sets. The C algorithm and WinBUGS yielded the same results, but the C algorithm is orders of magnitude faster than the WinBUGS implementation, capable of analyzing a dataset in under 30 seconds on a standard modern PC.

of the slopes, resulting in inefficient estimates of the group means. However, we made the range of the prior distribution sufficiently large to allow for extreme heterogeneity across groups in growth rates so that the prior will not unduly influence the estimates. For example, in analyses of 19 outcomes for racial/ethnic groups using data from the NHIS, we found that the standard deviation of the slopes never exceeded $0.02r$.[5]. When bayesmodel =common, the prior for the common slope has a mean of $c_7 = 0$ and variance $c_8 = 1000000r^2$, so the prior dsitribution has essentially no influence on the estimate of the slope.

The $N(c_9 = 0, c_{10} = 1)$ prior for ψ, the transformation of ρ, leads to a prior distribution for ρ that has prior mode near zero and puts reasonably high prior mass on all but extreme values of ρ very near ± 1. Finally, we use a $U(c_{11} = 0.0001, c_{12} = 0.1r)$ prior distribution for τ. As with our prior on the standard deviation among group-specific slopes, we set the range of the prior to be sufficiently large to avoid undue influence on the parameter estimates. For example, the upper bound of the range $0.1r$ allows for much greater variability around the trend than we observed for any of the 19 outcomes included in our analysis of the NHIS data. To prevent convergence problems with the algorithm, which occur when $\tau \to 0$, the lower bound for the prior is not zero (the lower limit for an unrestricted range for a standard deviation), and it is not a function of r. The lower-bound value of τ, however, is sufficiently small to permit inferences about η_{gt} to be negligibly different from what would be obtained when $\tau = 0$, and so enforcing the lower bound to be just above zero has no practical consequences for the main inferences. Expert users could modify the bounds of the uniform priors using macro options available in Section 7.1.

5.5 Maximum Likelihood Estimation and Model-Averaging

The above procedures use Bayesian methods to estimate the statistical model parameters used by the MKF procedures to produce population mean estimates. The MKF can also use maximum likelihood estimates of the model parameters. In this approach, the parameter values for the intercepts and slopes of the time trends and the parameters of the AR(1) process for the deviations from the trend are chosen to maximize the likelihood of the observed data given the specified statistical model. These parameters are then used to deterministically calculate the predicted population mean using the iterative Kalman Filtering algorithm (Elliot et al., 2009; Kalman, 1960).

As with the Bayesian approach, the MKF macro allows users to supply user-defined restrictions on the slope parameters for the group time trends to potentially improve on the precision of the estimates. The user can allow each group to have a separate unspecified slope (slopes =independent), for all groups to share a single common slope (slopes =common), or for no time trend in any group (slopes =dropped). The parameters

[5] See Elliott et al. (2009) for details on the outcomes and the definitions of racial/ethnic groups

of each model are then estimated via the PROC NLMIXED procedure in SAS©.

Research has found that with few time points, the estimation of separate time trends for each group is the limiting factor in the gains in accuracy available from the MKF procedure (Elliott et al., 2009). If the user knew that trends were common or nonexistent across groups, then using this knowledge to specify restricted models could result in MKF estimates with superior gains in efficiency over an MKF procedure that estimated separate slopes for every group. However, if the user's assumption about a common or nonexistent slope is wrong, the accuracy of the MKF can be severely degraded by bias in the estimates of the slopes due to inappropriate model restrictions.

Simulation studies demonstrated that model-averaging provided a method for recouping most of the gains in accuracy from restricting the slopes when the slopes were truly common or zero or had very small variability across groups but without incurring the cost from severe model misspecification from restricting the slopes in settings when the slopes varied considerably across groups. The model average estimate equals a weighted average of the predicted population means estimated using a different MKF specification. The weighted average depends on the the relative sizes of the Bayesian information criteria (BIC) model fit indices for each of the different model specifications with the weights being a function of the BIC. In particular with J alternative estimation methods (e.g., $J = 3$ for models with separate unspecified slopes, a common slope, and no slope for the groups, slopes =independent common dropped or $J = 2$ for models with separate unspecified slopes and a common slope for the groups, slopes =independent common), the relative fit of the model $j = 1,\ldots,J$ to the last model is

$$\Delta BIC_{jJ} = -2l_j + 2l_J + (k_j - k_J)log(n),$$

where l_j is the log likelihood evaluated at the MLE for model j and k_j, $j,\ldots,J$ is the number of parameters estimated in model j, and n is the sample size. In applications without a by group, $n = GT$. The weight for model j is then given by

$$w_j = \frac{e^{-\frac{1}{2}\Delta BIC_{jJ}}}{\sum_{j'=1}^{J} e^{-\frac{1}{2}\Delta BIC_{j'J}}}.$$

The quantity $e^{-\frac{1}{2}\Delta BIC_{jJ}}$ approximately equals the Bayes factor for comparing the models (Kass and Raftery, 1995), and the weights are approximately the optimal weights for combining the estimates from the alternative models.

When the data demonstrate very little difference between the different groups in time trend, the model with a common slope (or no slope) will tend to fit the data well, have a relatively small value for BIC, and have greater weight than a model that allows for separate slopes. Thus, the model average estimate would capture the gains in accuracy from placing restrictions on the slopes. Alternatively, if the data demonstrate large variability in the trends among the groups, the model with the common slope (or no slope)

will tend to fit the data poorly, have a relatively large value of BIC, and receive lower weight. Thus, the model average protects against relying on restrictive model specifications when that is incorrect.

The full Bayesian model specification uses the prior variance in the group slopes in a similar fashion to adaptively determine how much weight to give to each individual group's data when estimating its slope. This allows the full Bayesian specification to make efficient use of the data without relying on potentially overly restrictive model assumptions. Simulation studies found that full Bayesian procedures (bayesmodel =full) and the MLE procedures with model averaging of models with separate slopes and a common slope (slopes =independent common) or with all three models (slopes =independent common dropped) yield very similar estimates.

The MLE procedures do not provide a straightforward method for estimating the RMSE of the population mean predictions using only the sample means and standard error because the predictions are complex nonlinear functions of the parameter's AR(1) process. The RMSE estimates for the MKF macro MLE estimates follow the common practice of treating these parameters as known quantities rather than estimates. This results in an underestimation of the RMSE. The Bayesian procedures *do not* make this assumption and provide accurate estimates of the RMSE. Given the similarity of the predictions for the population means from our full Bayesian and our model average MLE estimates and the superiority of the RMSE estimates of the fully Bayesian procedures, users are generally advised to use the fully Bayesian procedures. The MLE estimates are provided for users to explore modeling alternatives and test the sensitivity of results to the default prior specifications.

5.6 Standard Errors of Zero

The model underlying the MKF assumes that the population mean is estimated with error, and the estimation code is designed only for data in which this is the case. As noted above, in some instances with rare events and small samples, samples in some groups may have no variance, and the estimated standard error of the mean may be zero. However, the mean is still estimated with error, but the user lacks information for estimating the variability of these sampling errors.

The MKF software imputes an alternative positive estimate for the standard error of these means. If $SE_{Y_{gt}} = 0$, the software imputes a value of $SE_{Y_{gt}} = \frac{1}{\tilde{T}} \sum_{\tilde{t}=1}^{\tilde{T}} SE_{Y_{g\tilde{t}}}$, the average standard error in the group g over the time points $\tilde{t} = 1, 2, \ldots, \tilde{T}$ where $SE_{Y_{g\tilde{t}}} \neq 0$. The macro will print a warning message in the output if such imputation was done. In the case where within a group g, $SE_{Y_{gt}} = 0$ for all $t = 1, 2, \ldots, T$, then there is no information for imputing the value of $SE_{Y_{gt}}$ and the analysis will not be conducted. Experienced statistical users can impute the standard errors in ways different from the one proposed by the software macro and conduct sensitivity analysis for their data on this assumption

of imputing zero standard errors with the average standard error over time. For some data, it might be more meaningful to take the average standard only for the time point before and the time point after the zero standard error observation. Experience users can build such imputations into their data before input into the macro for estimation.

6 Appendix. Details for Linux Users

For the Linux© or Unix© operating systems, the executable Bayesian estimation programs is kflinux. The file needs to be saved in the directory with MKF_MACRO. It will need to be set to have an EXECUTION permission. To do so, in the directory where the file is saved, execute the command:

$$chmod \quad 777 \quad kflinux.$$

This will make the kflinux file EXECUTABLE.

7 Appendix. Advanced Macro Parameters

These are parameters with the default option set in the macro but can be set by users with a good knowledge of the Bayesian estimation method who want to control the estimation method. The parameters are:

rho = <default = (empty)>;

tausq = <default = (empty)>;

df = <default = 1000>;

seed = <default = 1235>;

mcmcburn = <default = 10000>;

mcmciter = <default = 50000>;

chains = <default = 3>;

tolerence = <default = 0.15>;

malpha = <default = (empty)>;

palpha = <default = (empty)>;

mbeta = <default = 0>;

pbeta = <default = (empty)>;

betal = <default = 0>;

betau = <default = (empty)>;

mrho = <default = 0>;

prho = <default = 1>;

taul = <default = 0.0001>;

tauu = <default = (empty)>;

propsds = <default = 0.75 0.2 2>;

deletecsv = <default = yes>.

Below are the descriptions of these additional parameters.

7.1 Advanced Additional Parameters' Details

The first three parameters apply only to MKF procedures using maximum likelihood estimation of the model parameters, i.e., when slopes is not (empty).

rho = <default = (empty)>
: user-specified value for the autocorrelation parameter ρ. If not given (default= empty), it will be estimated. This option applies only to MKF with the MLE procedures (slopes not (empty)).

tausq = <default = (empty)>
: user-specified value of the innovation parameter τ^2. If not given (default= empty), it will be estimated for the procedures with (slopes not (empty)).

df = <default = 1000>
: specifies the denominator degrees of freedom used in testing model parameters (not predicted means) when MKF uses maximum likelihood for model estimation.

The remaining parameters apply only to MKF procedures using Bayesian estimation of the model parameters, i.e., when bayesmodel is not (empty).

seed = <default = 1235>
: specifies a random number-generating seed that will be used in the Bayesian model.

mcmcburn = <default = 10000>
: specifies the number of burn-in in MCMC iterations. If the Bayesian model does not converge, increasing this value might lead to convergence.

mcmciter = <default = 50000>
: specifies the number of post-burn-in in MCMC iterations. If the Bayesian model does not converge, increasing this value might lead to convergence.

chains = <default = 3>
: specifies the number of chains to run for the Bayesian estimation. Default is three.

tolerance = <default = 0.15>
: specifies the chain convergence tolerance. The default is set at 0.15.

malpha = <default = (empty)>
: specifies the prior mean for the group intercepts α_g, $g = 1, 2, \ldots, G$.

palpha = <default = (empty)>
: specifies the prior precision for the group intercepts α_g, $g = 1, 2, \ldots, G$.

mbeta = <default = 0>
specifies the prior mean for the mean slopes β_g across groups.

pbeta = <default = (empty)>
specifies the prior precision for the mean slopes β_g across groups.

betal = <default = 0>
specifies the lower bounds for *U(a,b)* prior for the standard deviation of slopes across groups.

betau = <default = (empty)>
specifies the upper bounds for *U(a,b)* prior for the standard deviation of slopes across groups.

mrho = <default = 0>
specifies the prior mean for transformed ρ.

prho = <default = 1>
specifies the prior precision for transformed ρ.

taul = <default = 0.0001>
specifies the lower bounds for *U(a,b)* prior for τ (standard deviation of innovation variance).

tauu = <default = (empty)>
specifies the upper bounds for *U(a,b)* prior for τ (standard deviation of innovation variance).

propsds = <default = 0.75 0.2 2>
specifies a vector of 3 Metropolis-Hastings algorithm proposal standard deviations $(\rho, \tau^2, var(\beta))$.

deletecsv = <default = yes>
If yes, the comma-separated values (csv) data file created from the C-program will be deleted, otherwise it will be saved in the working directiory. The csv files contain the MCMC draws from the model run and advanced users wishing to the MCMC draws can do so; for details they can contact the authors.

8 Acknowledgments:

This study was supported by contract HHSP23320095649WC, task order HHSA23337003T from the Office of Minority Health (HHS).

We thank Mary Roary, Garth Graham, Mirtha Beadle, and Rochelle Rollins of the Office of Minority Health for their support of this project. We also thank the reviewers James Hodges and Susan Paddock for their thorough reviews that greatly helped enhanced the quality and clarity of the manual. The RAND Health Quality Assurance process employs peer reviewers, including at least one reviewer who is external to the RAND Corporation. This study benefited from the rigorous technical reviews of the two reviewers (James Hodges and Susan Paddock), which served to improve the quality of this report.

9 Bibliography

Carlin, B.P. and Louis, T.A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Chapman and Hall/CRC Press.

Elliott M.N., McCaffrey D.F., Finch B.K., Klein D.J., Orr N., Beckett M.N., Lurie N. (2009). "Improving Disparity Estimates for Rare Racial/Ethnic Groups with Trend Estimation and Kalman Filtering: An Application to the National Health Interview Survey". *Health Services Research*, 44 (5), 1622-1639.

Gelman, A.; Carlin, J.B.; Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*, Chapman and Hall, London.

Gilks, W. R. and Richardson S. and Spiegelhalter D.J. (1996). *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London.

Kalman, R. 1960. "A New Approach to Linear Filtering and Prediction Problems". *Transactions of the ASME-Journal of Basic Engineering* 82, 35-45.

Kass, R.E. and Raftery, A. (1995). "Bayes Factors". *Joumal of the American Statistical Association*, 90, 773-795.

Klein RJ, Proctor SE, Boudreault MA, Turczyn KM. (2002) Healthy People 2010 Criteria for Data Suppression. *Statistical Notes*, no 24. Hyattsville, Maryland: National Center for Health Statistics.

Lockwood J.R., Setodji C.M., Elliott M.N., and McCaffrey D.F. (2009). "Smoothing Across Time in Repeated Cross-Sectional Data". *CDC SAG Twelfth Biennial Symposium on Statistical Methods*.

Lockwood J.R., McCaffrey D.F., Setodji C.M. and Elliott M.N. (2011). "Smoothing Across Time in Repeated Cross-Sectional Data". *Statistics in Medicine* 30(5): 584-594.

Miller ST, Schlundt DG, Larson C, Reid R, Pichert JW, Hargreaves M, Brown A, McClellan L, and Marrs M (2004). "Exploring Ethnic Disparities in Diabetes, Diabetes care, and Lifestyle Behaviors: the Nashville REACH 2010 Community Baseline Survey". *Ethnicity and Disease* 14(3 Suppl 1):3845.

Setodji C.M., Adams J.L., McCaffrey D.F., Elliott M.N., and Roary M. (2011). "Borrowing Strength Across Time and Outcomes in Repeated Cross-Sectional Data". Manuscript under preparation.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). “WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility”. *Statistics and Computing* 10, 325-337.