NLSY79 data: smoking habits of youth

Contents

1. Introduction 1

2. Building the path diagram 4

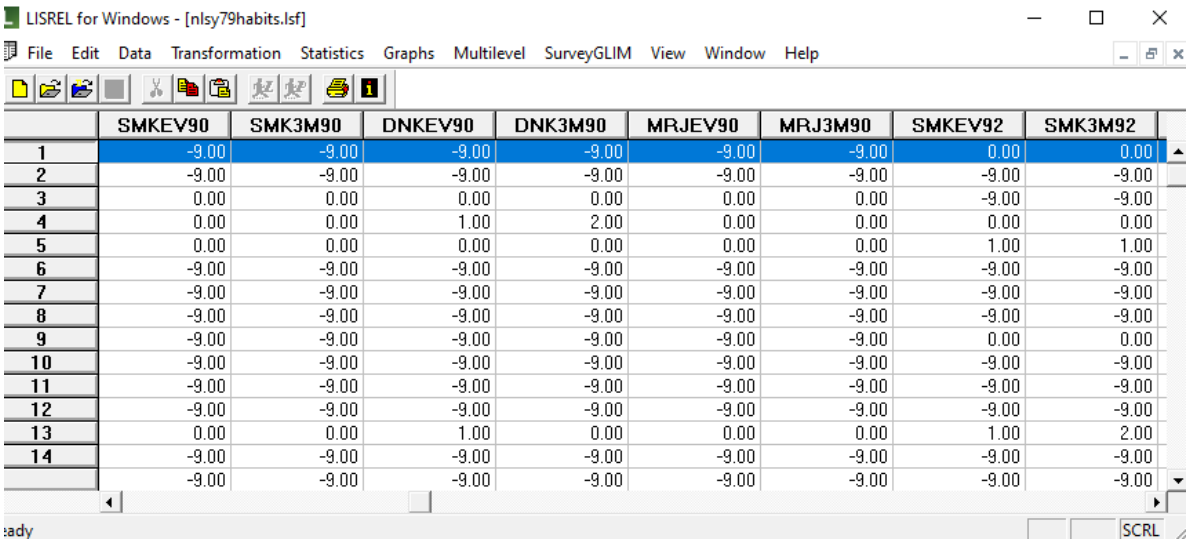
3. Generating syntax and running the model 11

1. Introduction

The data for the current study were obtained from the National Longitudinal Survey of Youth (NLSY) of Labor Market Experience in Youth, a study initiated in 1979 by the United States Department of Labor in order to study the transition of young people into the labor force. The NLSY is a multistage, clustered probability sample if households.

In 1986, a separate survey of all children born to NLSY79 female respondents began, greatly expanding the breadth of child-specific information collected. In addition to all the mother's additional demographic and development information collected from either the mother or child. For children aged 10 or older, information has been collected from the children biennially since 1988 on a variety of factors including attitudes towards schooling, dating and friendship patterns, health, substance use, and home responsibilities.

In this example we use a subset of these data (**nlsy79habits.lsf**) and look specifically at questions relating to smoking. The first few lines of the data are shown below. This data set contains information on smoking, drinking and the use of marijuana.



	SMKEV90	SMK3M90	DNKEV90	DNK3M90	MRJEV90	MRJ3M90	SMKEV92	SMK3M92
1	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	0.00	0.00
2	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
3	0.00	0.00	0.00	0.00	0.00	0.00	-9.00	-9.00
4	0.00	0.00	1.00	2.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
6	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
7	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
8	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
9	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	0.00	0.00
10	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
11	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
12	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
13	0.00	0.00	1.00	0.00	0.00	0.00	1.00	2.00
14	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00
	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00	-9.00

A short description of the variables is as follows:

- MOMID: Mother's identification code
- MOMWT79: 1979 sampling weight for mother
- CHILDID: Child identification code
- CHAGE: Child's age in years at the first measurement occasion (1988)
- CRACE: Race of child (1 = white, 2 = black, 3 = non-black, non-hispanic)
- CSEX: Gender of child (1 = male, 2 = female)
- REGION: region of residence of the mother of the child (1 = North East, 2 = North Central, 3 = South, 4 = West)

The substance abuse variables SMKEV88, SMK3M88, …, MRJPRS96 listed above are distinguished by three letters:

- INCO88 – INCO96: Family income for the years 1988, 1990, …, 1996
- SMK = cigarette use
- DNK = alcohol use
- MRJ = marijuana use

For the substance use variables, the first three letters are followed by the subscripts

- EV : Ever used (0 = no, 1 = yes)
- 3M: used during last three months (0 = never, 1 = no, 2 = yes)
- PRS: pressured to use tobacco or other substances (0 = no, 1 = yes)
- 30D: used during last 30 days (0 = no, 1 = yes)
- REC: how recently used. Scale from 0 to 7 indicating never to very recently
- RACE1, RACE2 and RACE3 are dummy coded variables created from CRACE.

Of interest here is the smoking habits of youth over time. In 1988, 1990 and 1992 two questions on the use of tobacco were asked. In the last two completed questionnaires, 1994 and 1996, five questions were answered. These variables are assumed to measure the same underlying concept, namely cigarette use. For each of the 5 occasions, we assume that the use of the tobacco related questions are the dependent variables. In LISREL context, these are Y-variables. We assume that the five latent variables (cigr88, cigr90, cigr92, cigr94, and cigr96) are indicators of cigarette use on the five measurement occasions. When latent variables are associated with dependent variables, the latent variables are ETA variables in the LISREL framework.

We also have demographic information on gender, age and race that can be used as predictors of the Y-variables. These predictor variables are X-variables in the LISREL framework, and we include them in the model as potential predictors of the latent variables. While it is not necessary for the user to memorize the eight matrices traditionally used in LISREL, the above definitions are important as they play a role in correctly constructing a model by building a path diagram.

As the focus of this analysis is on tobacco use only, we use a reduced LSF file containing only demographic and tobacco usage related variables. The data set used in this example is **NLSY79cigs.LSF** and is obtained by running the following PRELIS syntax file (**nlsy79habits.prl**):

```
SY='nlsy79habits.LSF'
SE 5 11 12 17 18 23 24 31 32 33 34 35 43 44 45 46 47 60 61 62
```

```
OU RA=nlsy79habits.lsf XM
```

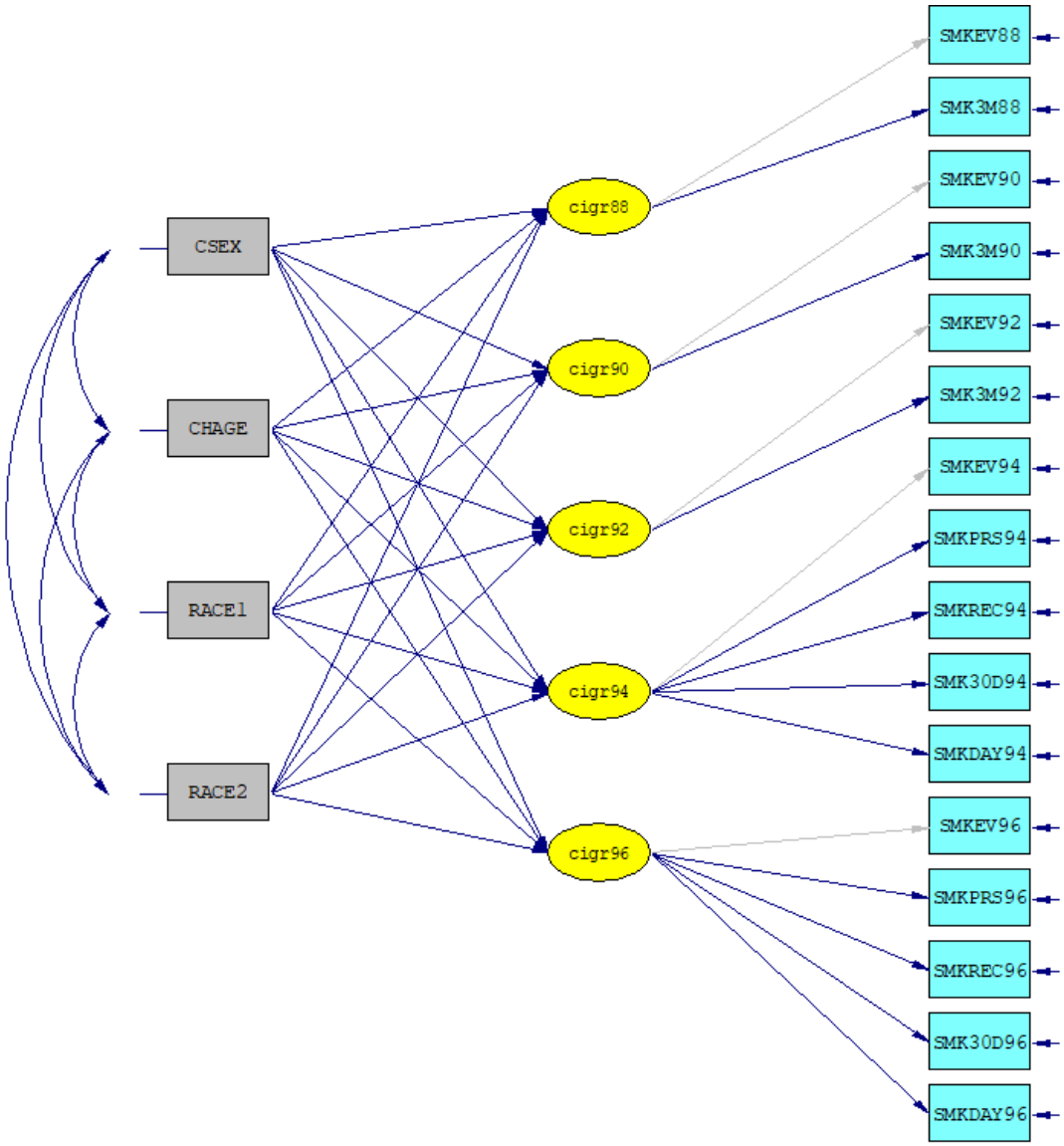
The first few lines of the LSF are shown below. Missing data values are denoted by -999999. In order to use all the data available, we will use Full Information Maximum Likelihood (FIML) estimation.

LISREL for Windows - [NSLY79cigs.LSF]

File Edit Data Transformation Statistics Graphs Multilevel SurveyGLIM View Window Help

	CHILDID	SMKEV88	SMK3M88	SMKEV90	SMK3M90	SMKEV92	SMK3M92	SMKEV94
1	301.0	-999999.0	-999999.0	-999999.0	-999999.0	0.0	0.0	0.0
2	302.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	0.0
3	401.0	-999999.0	-999999.0	0.0	0.0	-999999.0	-999999.0	-999999.0
4	801.0	0.0	0.0	0.0	0.0	0.0	0.0	-999999.0
5	802.0	-999999.0	-999999.0	0.0	0.0	1.0	1.0	-999999.0
6	803.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	0.0
7	1001.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	0.0
8	4301.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	0.0
9	4401.0	-999999.0	-999999.0	-999999.0	-999999.0	0.0	0.0	-999999.0
10	4402.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0	-999999.0

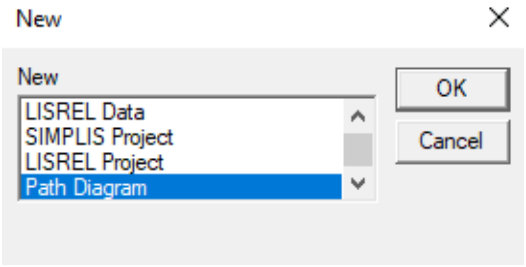
Ready



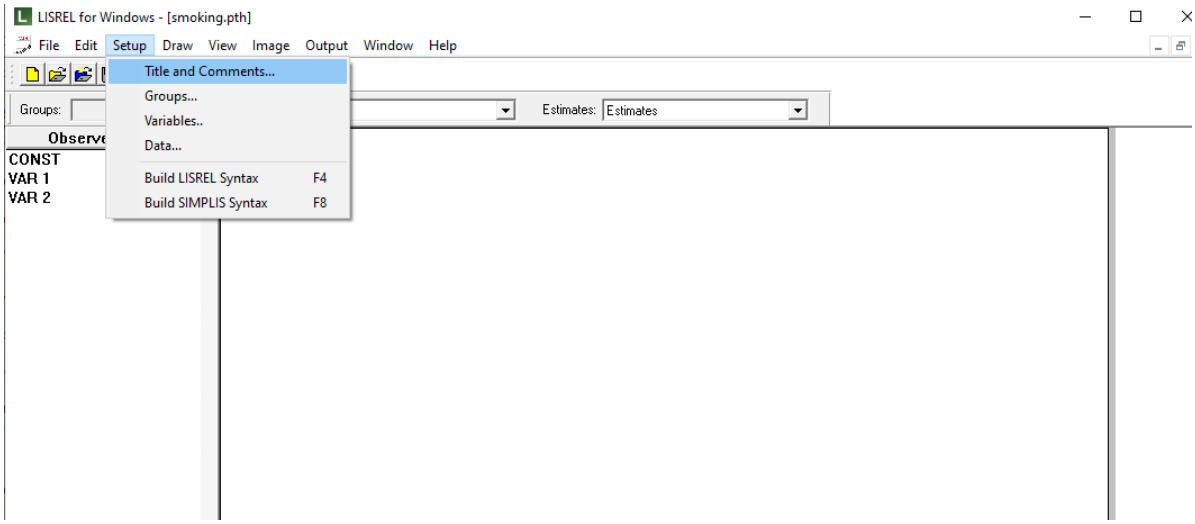
A conceptual path diagram of the model we would like to fit is shown above. Note that there is a color difference between paths originating from the latent variables and the dependent variables: some appear grey, others blue. The reason for this is that each grey path has been fixed to a value of 1.0. The reason for this is that one path between a latent variable and a dependent variable must be fixed to 1. Latent variables have no scale or location, so fixing one path allows the other path(s) to be interpreted relative to the fixed path.

2. Building the path diagram

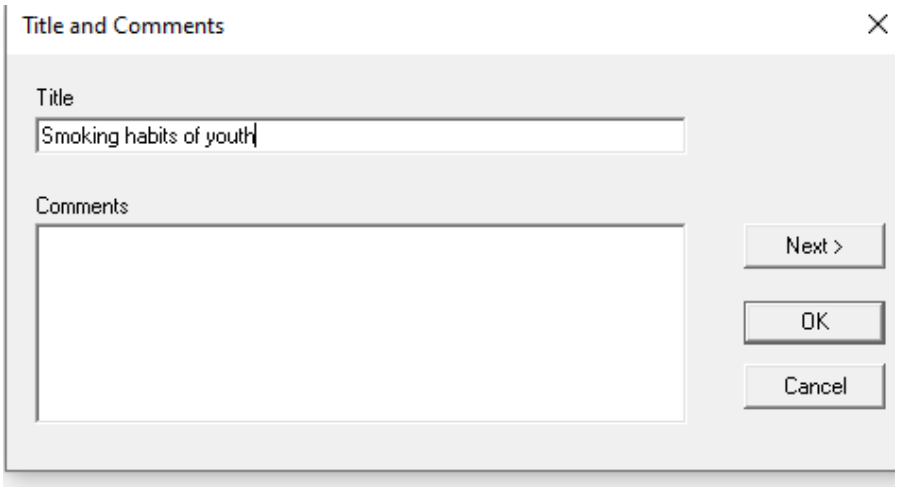
As a first step, click on **File**, **New** to create a path diagram. Select the Path Diagram option as shown below and click **OK**. Assign a name to the path diagram when prompted to do so.



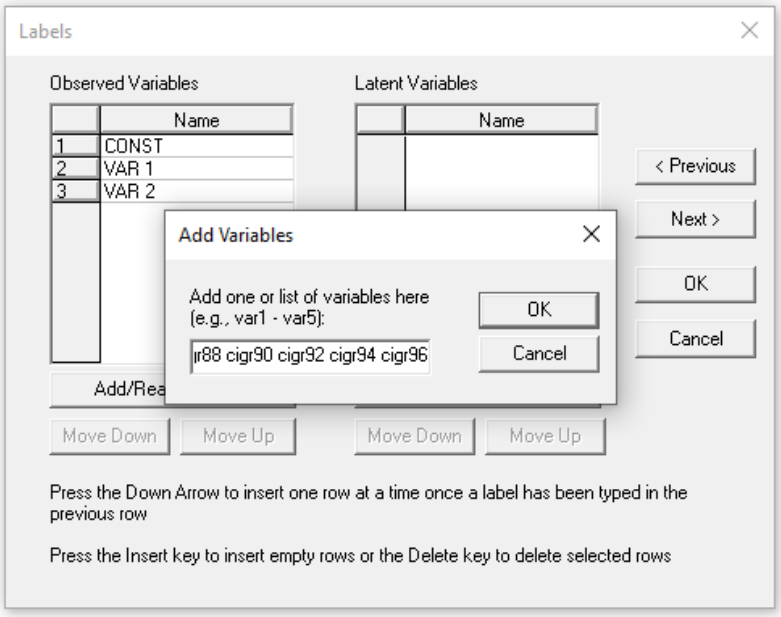
The path diagram window is now displayed. The **Setup** menu is used to provide information on the model. Start by selecting **Title and Comments** from this menu.



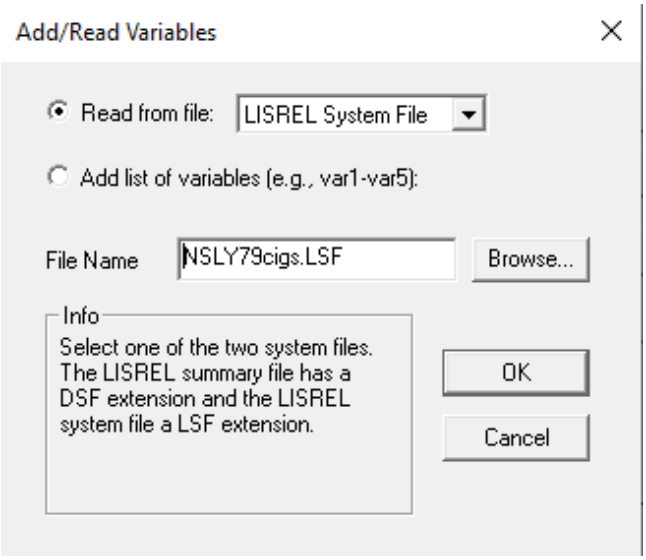
On the **Title and Comments** dialog box, enter a title and any comments, and click **Next** to proceed to the **Group Names** dialog box. As this is a single group analysis, we do not enter any information here but simply click next to go to the **Labels** dialog box. Note that this dialog box is associated with the **Variables** item on the **Setup** menu.



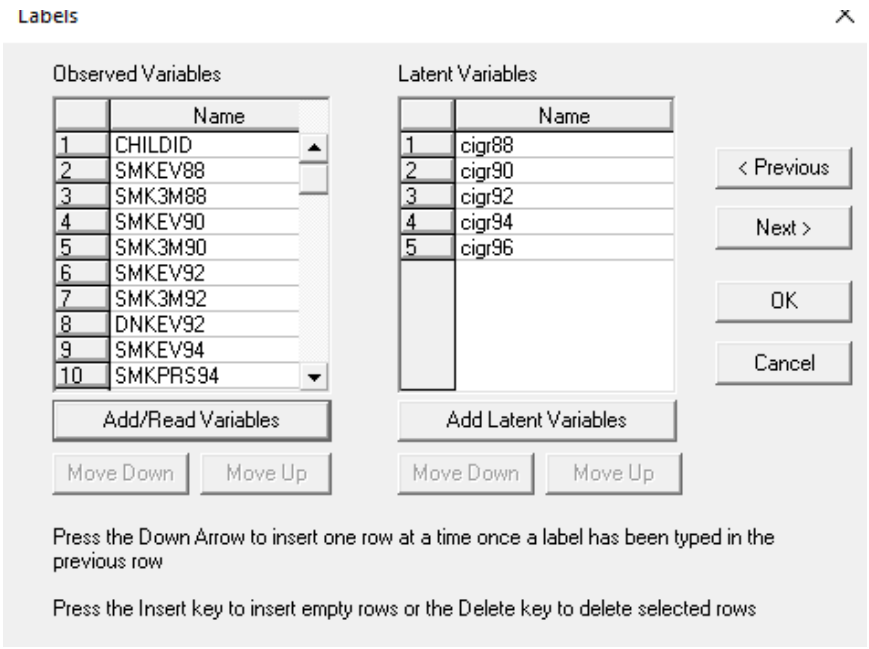
As we want to use 5 latent variables, one each for years 1988, 1990, 1992, 1994, and 1996, we click the **Add Variables** button below the **Latent Variables** field and add these five names to this field. Click **OK** when done.



The observed variables are read from the data file **nlsy79cigs.lsf**. To read in the data, first click on the **Add/Read Variables** button below the **Observed Variables** field. On the **Add/Read Variables** dialog box, select LISREL system file from the **Read from file** drop-down list. **Browse** for and select the LSF file named **nlsy79habits.lsf** as shown below before clicking **OK**.



Clicking **OK** on the **Add/Read Variables** dialog box returns you to the **Labels** dialog box, where the variables in the LSF are now displayed in the **Observed Variables** field. As all variables are now available, click **OK**.



On return to the path diagram window, all the variables are now displayed at the left.

We now proceed with the actual drawing of the path diagram. A good guide for easy path diagram construction is to do it in the following sequence:

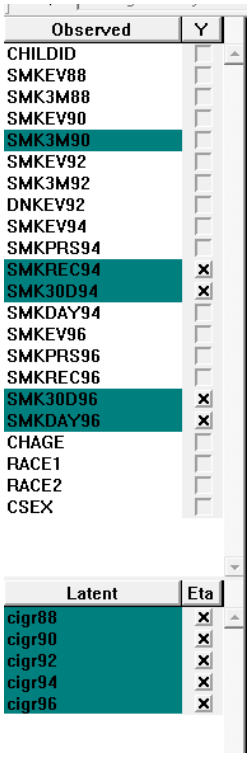
1. Select all the Y-variables by checking the box under the **Y** heading.
2. Indicate the Eta-variables by checking the box under the **Eta** heading.
3. Drag and drop dependent and independent variables into the Path diagram window, followed by the Ksi and Eta variables.
4. Add paths between the variables.
5. Fix paths as needed.

Note also that

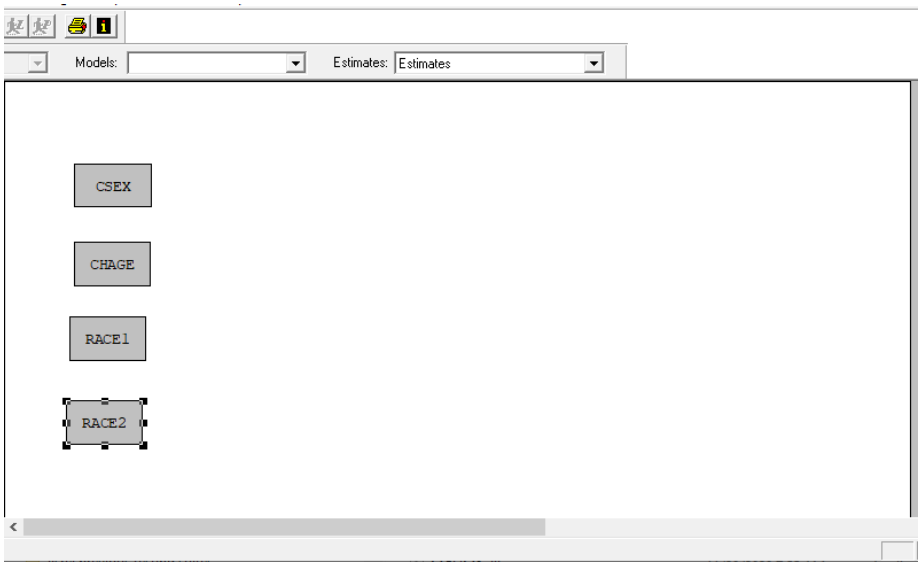
- The observed variables are assumed to be X (or independent) variables unless appropriate squares under the Y-column are clicked.
- The latent variables are assumed to be KSI (or independent) variables unless appropriate squares under the Eta-column are clicked.

As a first step, indicate that all the variables with a "SMK" prefix in the data are Y-variables by clicking the boxes under **Y** in the **Observed** field as shown below.

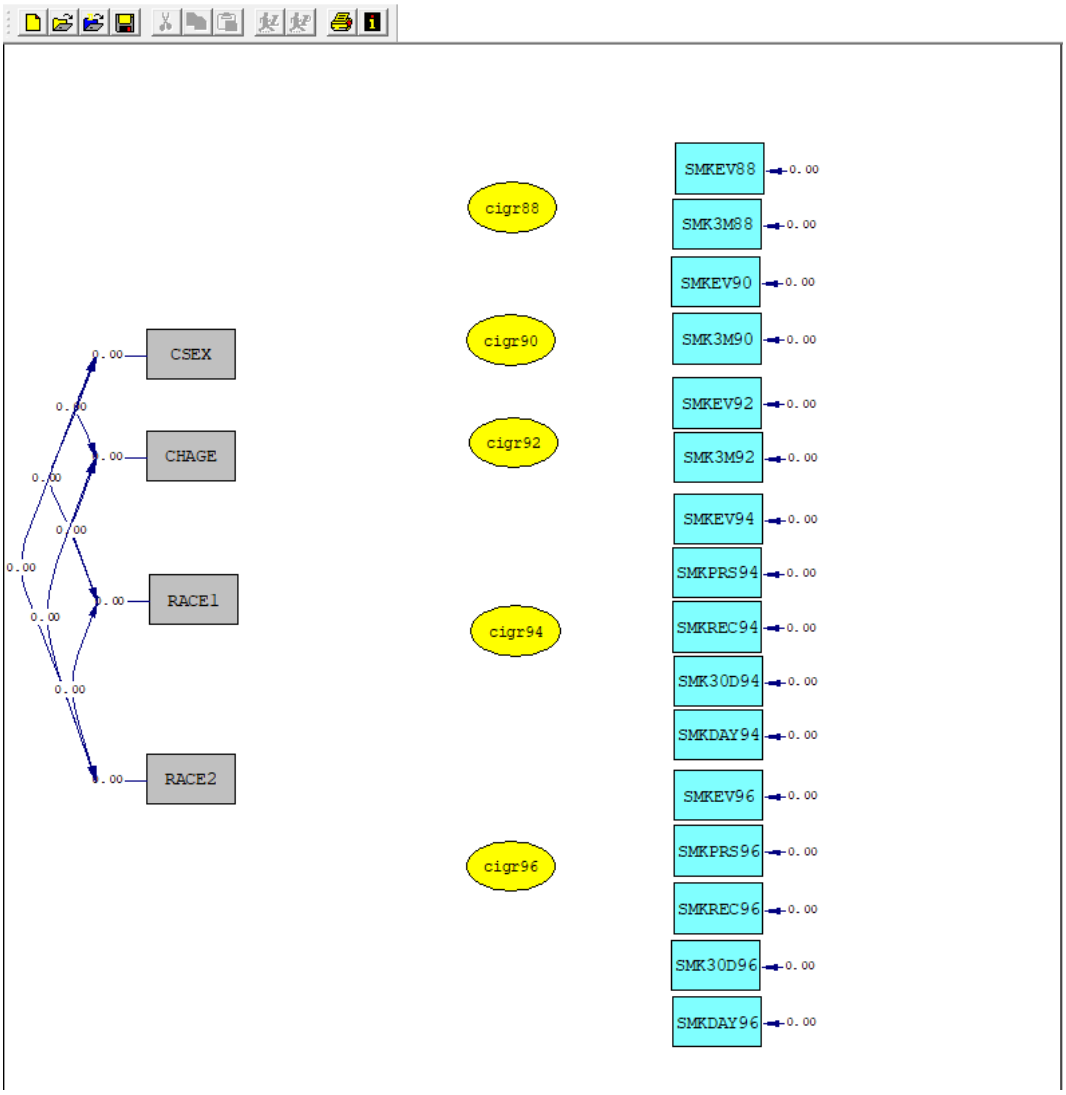
In this case, the latent variables are Eta variables, so check the boxes in the Eta column for all five variables in the **Latent** field. The variable list with Y-variables and Eta-variables selected is shown below.



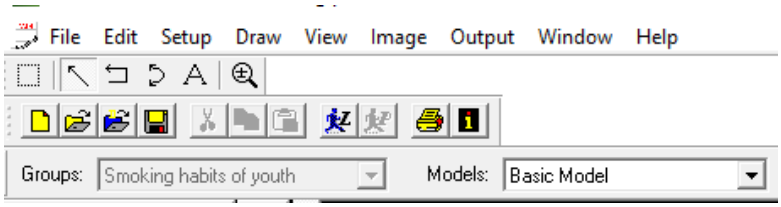
Start by clicking on the CSEX label under the **Observed** variables portion of the **Labels** window. Hold the mouse button down and "drag" the label to the draw area indicated by the grid lines. A rectangular-shaped object will appear on this part of the screen when the mouse button is released as shown below.



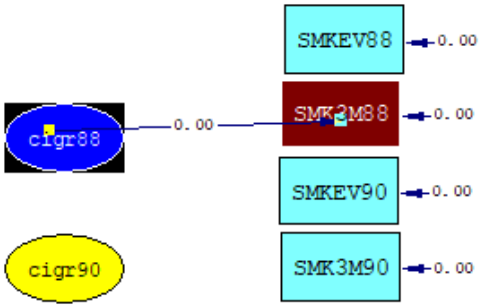
Once the four X-variables have been dragged to the screen, do the same for the five latent variables. Place the latent variables in the middle of the window. Finally, add the Y-variables at the right of the window. Once all variables have been entered, the path diagram will look as shown below.



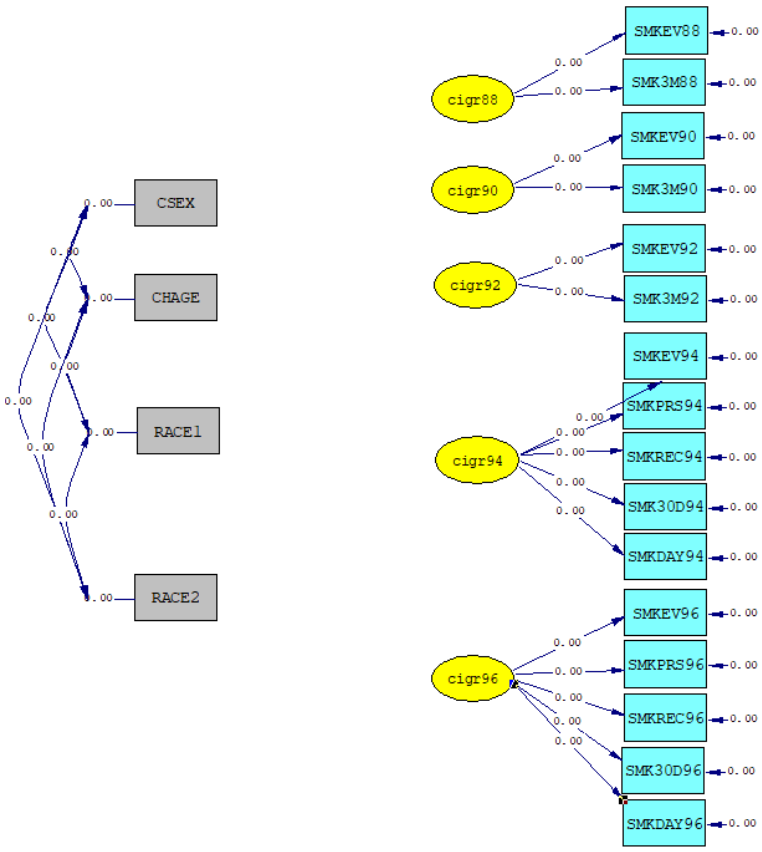
Next, we add the paths between the latent variables and the Y-variables. This is done by either using the floating toolbox or, depending on your selection, the Toolbox toolbar displayed at the top on the window. Display of this is set via the **View**, **Toolbars** menu.



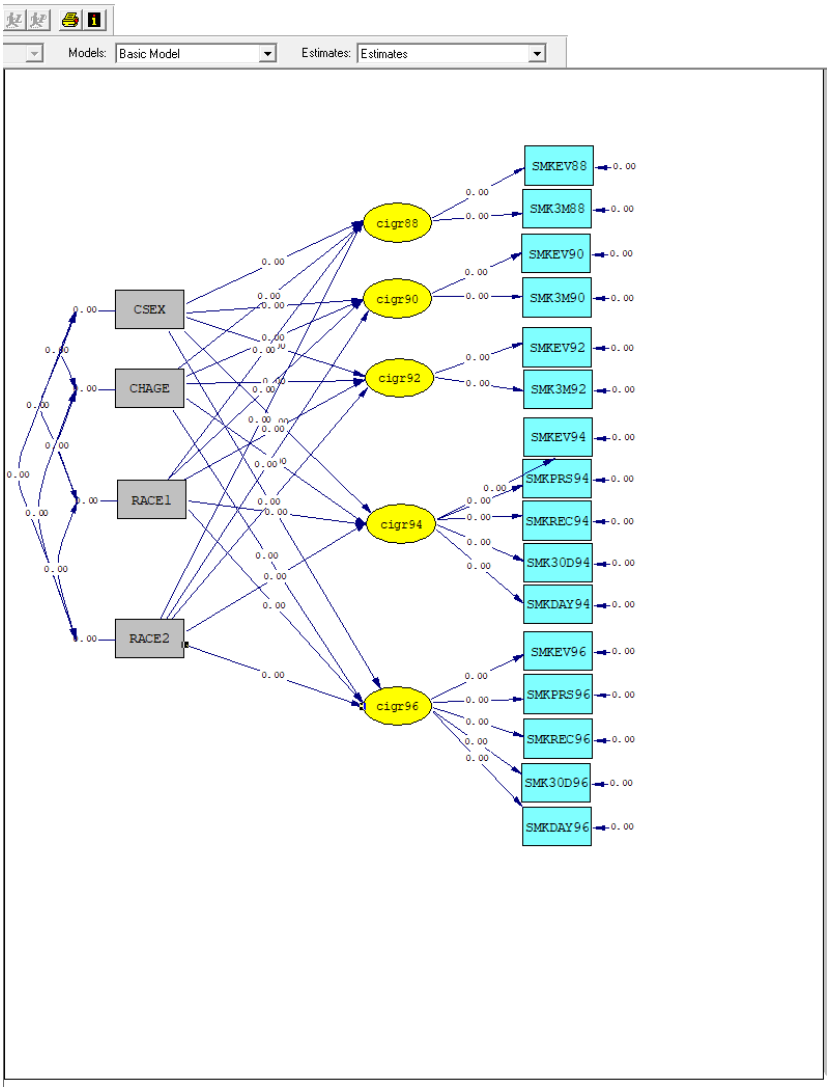
Click on the one-way arrow and draw paths from within a latent variable element to the Y-variables from the same year. With the left mouse button held down, drag the arrow to within the rectangular-shaped object. Release the mouse button when the colors of both objects change. For each latent variable, draw paths from within that latent variable element to the Y-variables from the same year. In the image below, a path is being added between cigr88 and SMK3M88.



When complete, the path diagram should correspond to the one shown below.

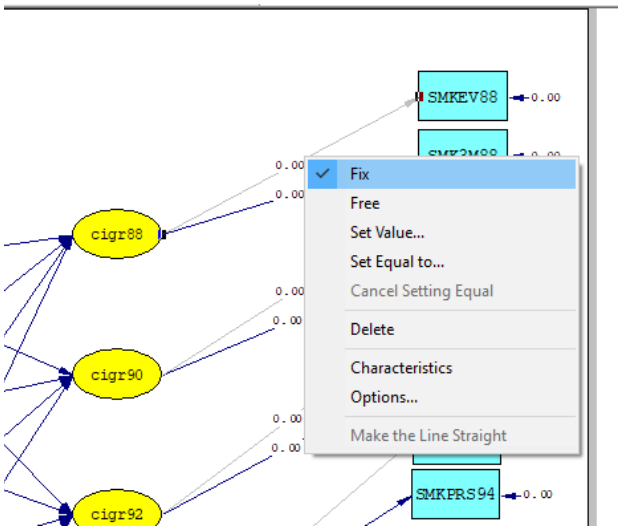


We now turn to adding the paths between the X-variables and the latent variables. To add these, select the one-directional arrow again. Start within each of the X-variables rectangular representations and draw a path to each of the latent variables. Release the mouse once the path has been drawn to within the ellipse representing the latent variable. After completion, the path diagram window should look as shown below.



Recall from Section 1 that some of the paths between the latent and Y-variables must be fixed, while others are free. Before generating syntax, we also fix these paths. Specifically, the paths between cigr88 and SMKEV88, cigr90 and SMKEV90, cigr92 and SMKEV92, cigr94 and SMKEV94, and cigr96 and SMKEV96 must be fixed.

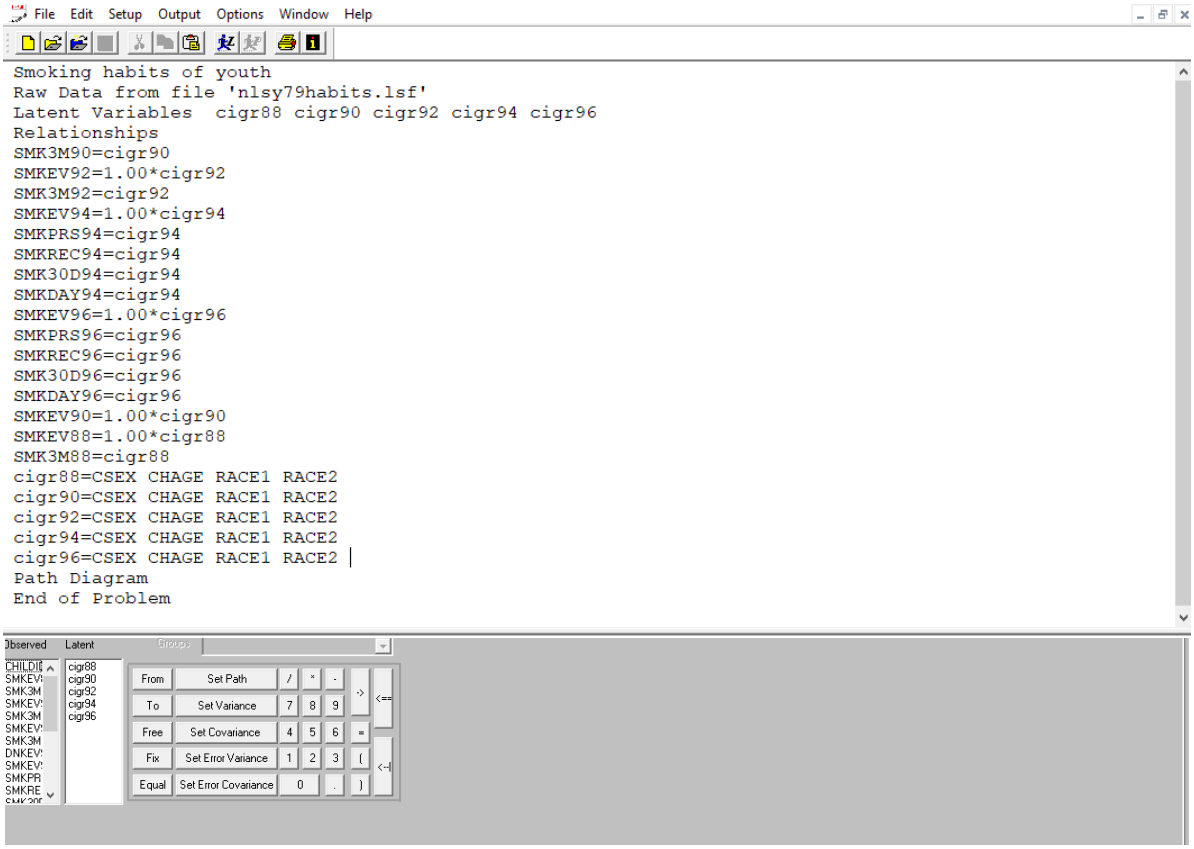
This is done by first selecting the **Select** tool from the toolbar (the rectangle). Next, right-click on the first path and select the **Fix** option from the pop-up menu.



Right-click again and select the **Set Value** option. On the **Set Starting or Fixed Values** dialog box, enter the value 1.0 and click **OK**. Repeat this for the other fixed paths.

3. Generating syntax and running the model

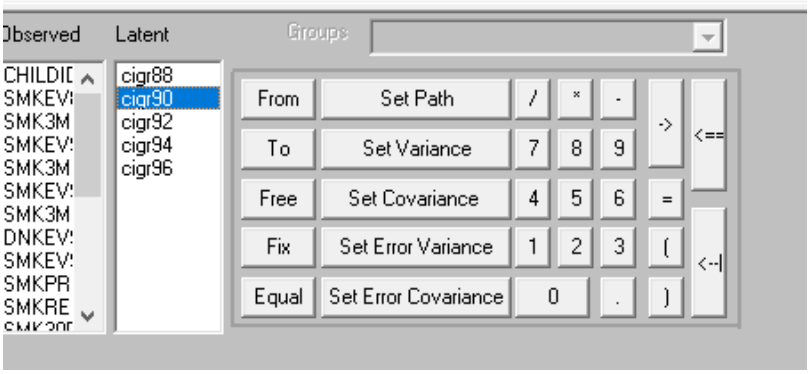
We are now ready to generate syntax for our model. Select **Build SIMPLIS syntax** from the **Setup** menu. We obtain the following SIMPLIS project file. (**cigarettes.spl**)



We now use the pad at the bottom of this window to set all the covariances between the latent variables free.

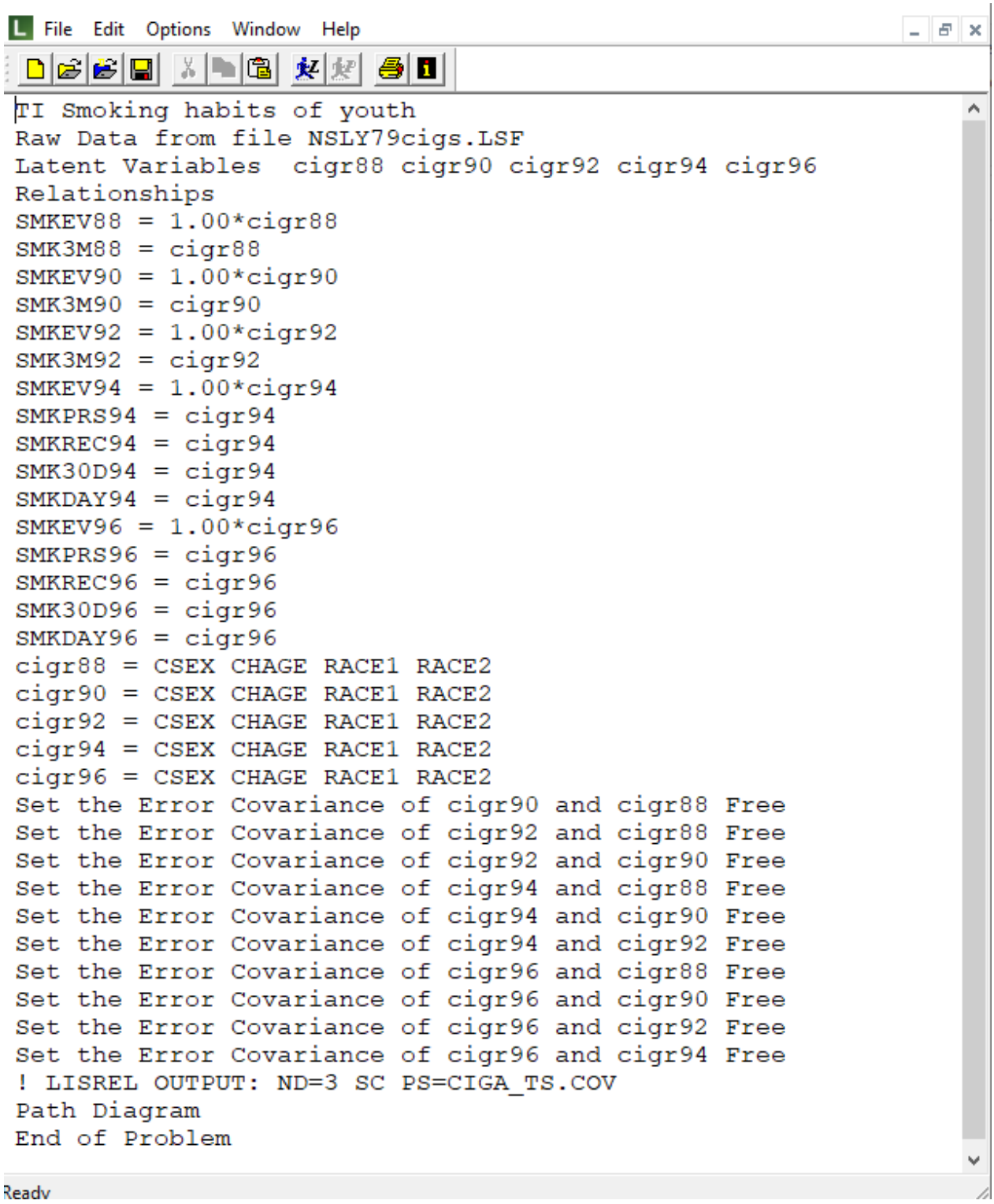
First, move to just before the Path Diagram line and press the Enter key to add a blank line. Next, click the **Set Error Covariance** button. The text **Set Error Covariance of** is now written to the previously empty line. Click on the latent variable cigr90 and drag it next to the text.

```
cigr92=SMK3M90
cigr94=SMK3M90
cigr96=SMK3M90
Set the Error Covariance of cigr90
Path Diagram
End of Problem
```

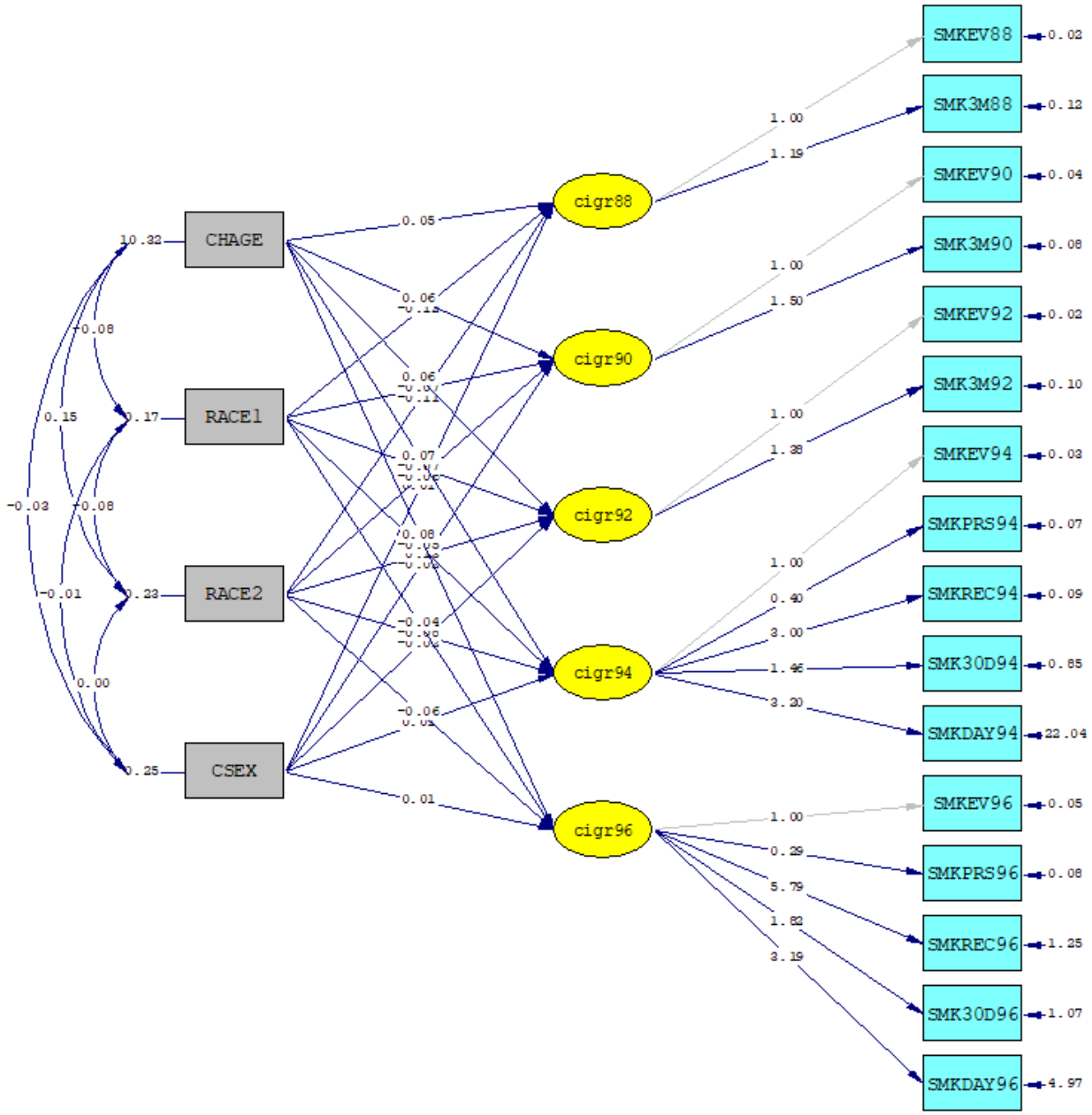


Click **And** (or simply type that in) and drag cigr88 to the line, followed by either clicking the **Free** button or typing the word Free at the end of the line to obtain the completed line as shown below. Repeat these actions for all error covariances between all five latent variables.

The final syntax file looks like this:



We can now run the model to obtain the following path diagram.



Chi-Square=468.00, df=138, P-value=0.00000, RMSEA=0.025

The first section of the output file contains the measurement equations. Judging by the R^2 for the individual equations, the dependent variables indicating whether a respondent has ever smoked shows a strong relationship with the corresponding latent variable. The same is true, to a lesser extent, for the cigarette use in the last three months, only available in 1988, 1990 and 1992. The R^2 for the measurement equations for SMKPRS94, SMK30D94, SMKDAY94, SMKPRS96, SMK30D96, and SMKDAY96 are noticeably smaller, though effects are still highly significant.

Measurement Equations

```
 SMKEV88 = 1.000*cigr88, Errorvar.= 0.0159  , R² = 0.895
Standerr                          (0.00683)
Z-values                           2.328
P-values                           0.020

 SMK3M88 = 1.192*cigr88, Errorvar.= 0.116  , R² = 0.625
Standerr  (0.0689)                (0.0113)
Z-values   17.314                  10.247
P-values   0.000                   0.000

 SMKEV90 = 1.000*cigr90, Errorvar.= 0.0371  , R² = 0.777
Standerr                          (0.00399)
Z-values                           9.304
P-values                           0.000

 SMK3M90 = 1.503*cigr90, Errorvar.= 0.0850  , R² = 0.774
Standerr  (0.0534)                (0.00904)
Z-values   28.127                  9.398
P-values   0.000                   0.000

 SMKEV92 = 1.000*cigr92, Errorvar.= 0.0188  , R² = 0.891
Standerr                          (0.00369)
Z-values                           5.104
P-values                           0.000

 SMK3M92 = 1.383*cigr92, Errorvar.= 0.0964  , R² = 0.754
Standerr  (0.0389)                (0.00770)
Z-values   35.563                  12.525
P-values   0.000                   0.000

 SMKEV94 = 1.000*cigr94, Errorvar.= 0.0334  , R² = 0.802
Standerr                          (0.00219)
Z-values                           15.236
P-values                           0.000

SMKPRS94 = 0.397*cigr94, Errorvar.= 0.0744  , R² = 0.223
Standerr  (0.0195)                (0.00264)
Z-values   20.366                  28.235
P-values   0.000                   0.000

SMKREC94 = 3.002*cigr94, Errorvar.= 0.0876 , R² = 0.933
Standerr  (0.0582)                (0.0169)
Z-values   51.554                  5.193
P-values   0.000                   0.000

SMK30D94 = 1.461*cigr94, Errorvar.= 0.848  , R² = 0.254
Standerr  (0.116)                 (0.0539)
Z-values   12.581                  15.730
P-values   0.000                   0.000

SMKDAY94 = 3.203*cigr94, Errorvar.= 22.039, R² = 0.0593
Standerr  (0.906)                 (2.141)
Z-values   3.535                   10.295
```

```
P-values   0.000                    0.000

 SMKEV96 = 1.000*cigr96, Errorvar.= 0.0503  , R² = 0.739
Standerr                           (0.00501)
Z-values                            10.047
P-values                            0.000

SMKPRS96 = 0.294*cigr96, Errorvar.= 0.0753  , R² = 0.141
Standerr  (0.0216)                 (0.00273)
Z-values   13.604                   27.572
P-values   0.000                    0.000

SMKREC96 = 5.787*cigr96, Errorvar.= 1.254 , R² = 0.792
Standerr  (0.250)                  (0.171)
Z-values   23.137                   7.339
P-values   0.000                    0.000

SMK30D96 = 1.816*cigr96, Errorvar.= 1.066  , R² = 0.306
Standerr  (0.156)                  (0.0807)
Z-values   11.615                   13.215
P-values   0.000                    0.000

SMKDAY96 = 3.191*cigr96, Errorvar.= 4.973 , R² = 0.226
Standerr  (0.458)                  (0.517)
Z-values   6.970                    9.624
P-values   0.000                    0.000
```

Turning our attention to the structural equations, we note that the gender of a respondent is not statistically significant at any of the five measurement occasions. This leads us to conclude that there is no significant difference in tobacco usage between males and females. Ethnicity, on the other hand, clearly plays a role, as estimates for the two ethnicity indicates are highly significant at the first four measurement occasions. On the last occasion, the estimate for RACE1 is not significant, but that for RACE2 is. Interestingly, the estimates associated with the respondent's age seem to be reasonably consistent in value over all occasions. When we look at the associated *z*-values, we note an increase over the first four measurement occasions, while the *z*-value at the last measurement occasions is slightly lower than that at the fourth occasion.

```
             Structural Equations


  cigr88 = 0.0506*CHAGE - 0.125*RACE1 - 0.114*RACE2 + 0.00972*CSEX, Errorvar.= 0.107   , R² = 0.214
Standerr  (0.00374)      (0.0316)      (0.0269)      (0.0238)                (0.00873)
Z-values   13.526        -3.955        -4.242         0.409                   12.220
P-values   0.000          0.000         0.000         0.682                   0.000

  cigr90 = 0.0621*CHAGE - 0.0715*RACE1 - 0.0581*RACE2 - 0.0191*CSEX, Errorvar.= 0.0886  , R² = 0.314
Standerr  (0.00317)      (0.0254)       (0.0216)       (0.0191)                (0.00560)
Z-values   19.624        -2.816         -2.685         -1.002                   15.835
P-values   0.000          0.005          0.007          0.316                   0.000

  cigr92 = 0.0594*CHAGE - 0.0684*RACE1 - 0.134*RACE2 - 0.0197*CSEX, Errorvar.= 0.116   , R² = 0.248
Standerr  (0.00271)      (0.0227)       (0.0194)      (0.0171)                (0.00551)
Z-values   21.935        -3.015         -6.944        -1.156                   21.110
P-values   0.000          0.003          0.000         0.248                   0.000

  cigr94 = 0.0713*CHAGE - 0.0512*RACE1 - 0.0827*RACE2 + 0.00881*CSEX, Errorvar.= 0.0828  , R² = 0.388
Standerr  (0.00249)      (0.0194)       (0.0165)       (0.0145)                (0.00371)
Z-values   28.637        -2.645         -5.004          0.606                   22.334
P-values   0.000          0.008          0.000          0.545                   0.000
```

```
  cigr96 = 0.0758*CHAGE - 0.0374*RACE1 - 0.0639*RACE2 + 0.00950*CSEX, Errorvar.= 0.0835  , R² = 0.414
Standerr  (0.00278)      (0.0228)       (0.0194)       (0.0171)                 (0.00583)
Z-values   27.263        -1.642         -3.290          0.554                    14.322
P-values   0.000          0.101          0.001          0.579                    0.000


Error Covariance for cigr90 and:cigr88 = 0.0370
                                        (0.00492)
                                           7.519

Error Covariance for cigr92 and:cigr88 = 0.0225
                                        (0.00568)
                                           3.961

Error Covariance for cigr92 and:cigr90 = 0.0414
                                        (0.00395)
                                          10.484

Error Covariance for cigr94 and:cigr88 = -0.004
                                         (0.0534)
                                          -0.066

Error Covariance for cigr94 and:cigr90 = 0.00697
                                         (0.00662)
                                            1.052

Error Covariance for cigr94 and:cigr92 = 0.0288
                                        (0.00373)
                                           7.709

Error Covariance for cigr96 and:cigr88 = 0.0812
                                         (0.0472)
                                           1.721

Error Covariance for cigr96 and:cigr90 = 0.0192
                                            (0.0)

Error Covariance for cigr96 and:cigr92 = 0.0158
                                        (0.00813)
                                           1.946

Error Covariance for cigr96 and:cigr94 = 0.0217
                                        (0.00356)
                                           6.099

  Covariance Matrix of Independent Variables

             CHAGE       RACE1       RACE2       CSEX
          ---------   ---------   ---------   ---------
   CHAGE     10.319
            (0.234)
             44.057

   RACE1     -0.075       0.168
            (0.021)     (0.004)
             -3.558      44.062
```

```
   RACE2      0.145     -0.079      0.233
             (0.025)    (0.003)    (0.005)
              5.813    -23.145     44.062

    CSEX     -0.029     -0.006      0.005      0.250
             (0.026)    (0.003)    (0.004)    (0.006)
             -1.127     -1.868      1.209     44.062
```

The estimated covariance matrix of the latent variables shown in the table below indicates higher covariances between the predictor CHAGE and all the latent variables than is observed for any of the other three predictors in the model. The covariances between the latent variables tend to decrease over the measurement occasions, except for the final measurement occasion. It may be that by the end of the period, the aging of the respondents over time may have led to respondents already having acquired a smoking habit or having opted to abstain.

```
         Covariance Matrix of Latent Variables

            cigr88     cigr90     cigr92     cigr94     cigr96      CHAGE
           --------   --------   --------   --------   --------   --------
  cigr88      0.136
  cigr90      0.071      0.129
  cigr92      0.055      0.080      0.155
  cigr94      0.035      0.053      0.073      0.135
  cigr96      0.121      0.068      0.062      0.077      0.143
   CHAGE      0.515      0.639      0.599      0.727      0.776     10.319
   RACE1     -0.016     -0.012     -0.005     -0.007     -0.007     -0.075
   RACE2     -0.009      0.001     -0.017     -0.005     -0.001      0.145
    CSEX      0.001     -0.006     -0.007      0.000      0.000     -0.029

         Covariance Matrix of Latent Variables

            RACE1      RACE2       CSEX
           --------   --------   --------
  RACE1      0.168
  RACE2     -0.079      0.233
   CSEX     -0.006      0.005      0.250

           Global Goodness of Fit Statistics, FIML case


             -2ln(L) for the saturated model =       51982.133
             -2ln(L) for the fitted model    =       52450.133

Degrees of Freedom = 138
Full Information ML Chi-Square                        467.999 (P = 0.0000)
Root Mean Square Error of Approximation (RMSEA)       0.0248
90 Percent Confidence Interval for RMSEA              (0.0224 ; 0.0273)
P-Value for Test of Close Fit (RMSEA < 0.05)          1.000
```

The modification indices indicate that there are changes that can be made to the current model in order to improve the fit. We leave it to the reader as an exercise.

The Modification Indices Suggest to Add the

Path to	from	Decrease in Chi-Square	New Estimate
SMKPRS94	cigr90	49.6	0.13

The Modification Indices Suggest to Add an Error Covariance

Between	and	Decrease in Chi-Square	New Estimate
SMK3M92	SMK3M88	61.2	-0.12
SMK3M92	SMK3M90	45.9	0.02
SMKEV94	SMK3M92	27.9	-0.01
SMKPRS94	SMKEV90	27.4	0.01
SMKPRS94	SMKEV92	10.9	-0.00
…			
RACE1	SMK30D94	9.5	-0.02
RACE1	SMKDAY94	14.3	0.11
RACE2	SMKDAY94	60.5	0.26