Stephen du Toit
Mathilda du Toit
Gerhard Mels
Yan Cheng

LISREL 使用手册

-PRELIS 应用范例

翻译：

程琰

温忠麟

目录

PRELIS 应用范例

PRELIS 是 LISREL 8.7 的一部分。它主要用于在构建结构模型之前，对数据进行前期处理和初步分析。PRELIS 的主要用途包括：

- 将其它格式的数据文件（SAS, SPSS, Excel, Stat 等等）读入并存储为 PRELIS 数据文件。
- 将 PRELIS（*.psf）数据输出为其他软件可读的相应格式。
- 对 PRELIS 数据进行处理。（定义变量类型，处理缺失值，数据筛选，生成子数据集等）。
- 回归模型分析及初步的因子分析等。
- 计算矩阵（协方差矩阵，多项相关系数以及渐近协方差矩阵等）。
- 可以用图表直观地表现数据的状况。

在这一章里，我们用不同类型的变量为例介绍如何应用 PRELIS 来实现上述功能。本章中的所有数据都存在 LISREL 安装文件夹中的 **TUTORIAL** 子文件夹里。

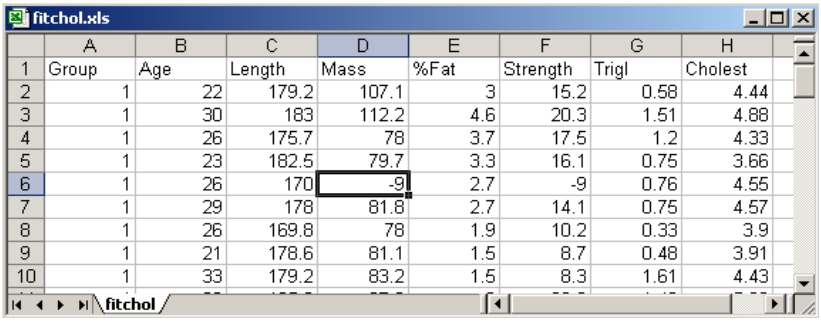
1.1 处理连续性数据

这一节里，我们主要介绍如何应用菜单选项对连续性变量进行处理。我们首先来示范如何用 PRELIS 进行数据清理。然后我们示范如何进行多元线性回归分析以及二阶最小二乘分析。

例1：以健身调查数据示范数据的清理和准备

关于数据

健身和血液中胆固醇的含量是影响心脏健康的两个重要因素。在一个相关的研究课题中，对三组成年男子，共 60 人进行了调查并记录了相关数据。这是一个 Excel 格式的数据集，其文件名是 **fitchol.xls**，存储在 **TUTORIAL** 子文件夹中。下表列出了前十个被调查者的观测数据。



fitchol.xls

	A	B	C	D	E	F	G	H
1	Group	Age	Length	Mass	%Fat	Strength	Trigl	Cholest
2	1	22	179.2	107.1	3	15.2	0.58	4.44
3	1	30	183	112.2	4.6	20.3	1.51	4.88
4	1	26	175.7	78	3.7	17.5	1.2	4.33
5	1	23	182.5	79.7	3.3	16.1	0.75	3.66
6	1	26	170	-9	2.7	-9	0.76	4.55
7	1	29	178	81.8	2.7	14.1	0.75	4.57
8	1	26	169.8	78	1.9	10.2	0.33	3.9
9	1	21	178.6	81.1	1.5	8.7	0.48	3.91
10	1	33	179.2	83.2	1.5	8.3	1.61	4.43

fitchol

注意：

-9 代表缺失值。在这个数据集中，共有三个观测对象的数据资料中出现缺失值。

这个数据集中的变量依次是：

- Group – 组别 (1 是举重运动员, 2 是学生，3 是长跑运动员) 。
- Age – 年龄。
- Length – 身高 cm。
- Mass – 体重 kg。
- %Fat – 脂肪百分比。
- Strength 肺活量 lb。
- Trigl – 甘油三酸脂肪。
- Cholest – 胆固醇。

要得到此数据集更详细的信息，请参考 Du Toit, Steyn 和 Stumpf (1986)。

数据准备和前期处理

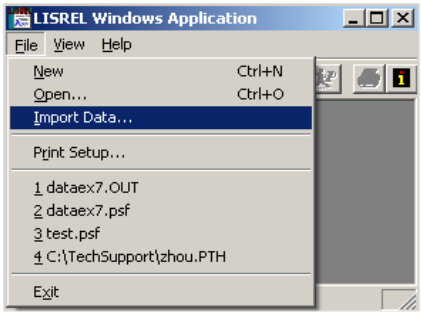
以数据集 **fitchol.xls** 为例，我们用图解说明怎样用 PRELIS 进行以下工作。

- 读取 Excel 数据
- 定义数据类型
- 定义整体缺失值 （global missing value）
- 插入新变量
- 给新变量赋值
- 选择观测对象并产生子数据集
- 将 PSF 数据文件转换为 Excel 文件

- 数据总览（data screening）
- 多元计算（multiple imputation）
- 计算正态值
- 计算矩阵

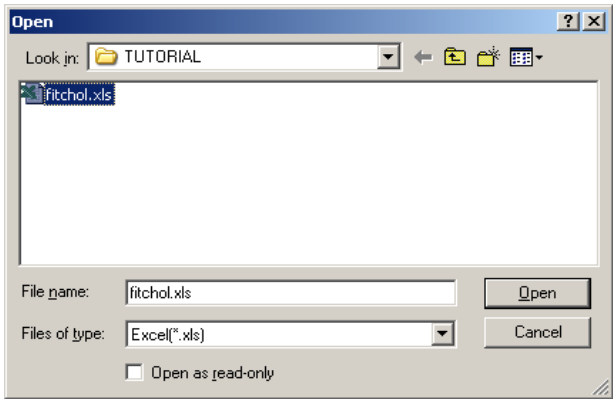
读入 Excel 数据文件

- 选择 **File** 菜单的 **Import Data** 选项

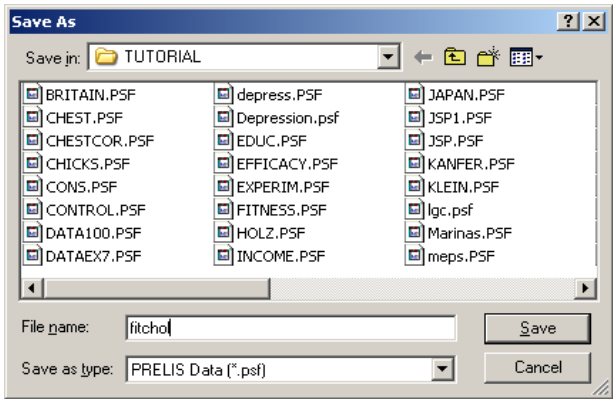


以打开 **Open** 对话框。

- 在 **Files of type** 下拉菜单中选择 **Excel (*.xls)** 选项。
- 如下图所示：找到并选择 **TUTORIAL** 子文件夹 **fitchol.xls** 文件。



- 点击 **Open** 键以打开 **Save As** 对话框。在 **File name** 字符区内输入 **fitchol** 得到如下对话框。



- 点击 **Save** 键以打开如下 PSF 窗口。

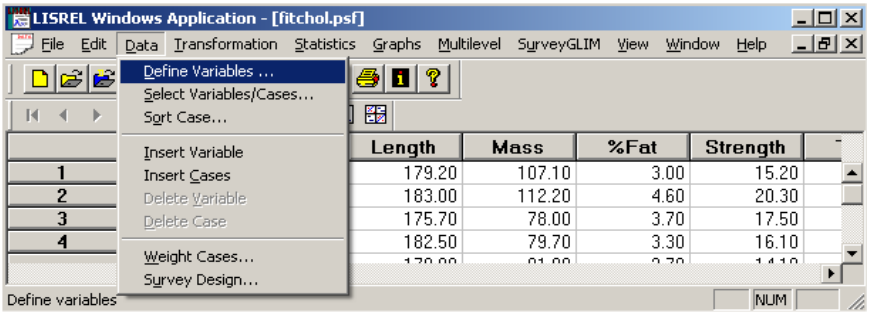
fitchol.psf

	Group	Age	Length	Mass	%Fat	Strength	Trigl	Cholest
1	1.00	22.00	179.20	107.10	3.00	15.20	0.58	4.44
2	1.00	30.00	183.00	112.20	4.60	20.30	1.51	4.88
3	1.00	26.00	175.70	78.00	3.70	17.50	1.20	4.33
4	1.00	23.00	182.50	79.70	3.30	16.10	0.75	3.66
5	1.00	26.00	170.00	-9.00	2.70	-9.00	0.76	4.55
6	1.00	29.00	178.00	81.80	2.70	14.10	0.75	4.57
7	1.00	26.00	169.80	78.00	1.90	10.20	0.33	3.90
8	1.00	21.00	178.60	81.10	1.50	8.70	0.48	3.91
9	1.00	33.00	179.20	83.20	1.50	8.30	1.61	4.43
10	1.00	36.00	185.20	87.80	6.00	23.80	1.42	5.33

定义变量类型

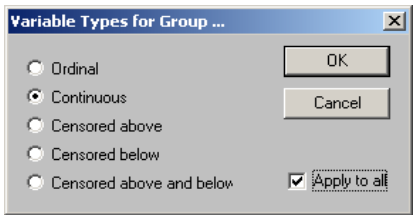
PRELIS 将 Excel 格式的数据文件读入并生成 PSF 文件时，变量类别默认为有序变量。在 **fictchol.psf** 这个数据集中，除了 Group 以外，其他变量都是连续的。如下，我们将示范如何从新定义变量类别。

- 点击 PSF 窗口中 **Data** 菜单的 **Define Variables** 选项激活 **Define Variables** 对话框。



- 从变量列表中选择变量 Group 激活 **Define Variables** 对话框上的所有键。

- 点击 **Variable Type** 键 打开 **Variable Types for Group…** 对话框。
- 激活 **Continuous** 选项按钮，选中 **Apply to all** 复选框得到 如下对话框。



- 点击 **OK** 键回到 **Define Variables** 对话框。
- 再点击 **Define Variables** 对话框上的 **OK** 键回到 PSF 窗口。
- 点击 **File** 菜单上的 **Save** 选项保存修改后的数据文件 **fictchol.psf**。
- 点击 **Data** 菜单上的 **Define Variables** 选项激活 **Define Variables** 对话框。
- 从变量列表中选择 变量 Group 。
- 点击 **Variable Type** 键 打开 **Variable Types for Group…** 对话框。
- 激活 **Ordinal** 选项按钮。
- 点击 **OK** 键回到 **Define Variables** 对话框。
- 再点击 **Define Variables** 对话框上的 **OK** 键回到 PSF 窗口。
- 点击 **File** 菜单上的 **Save** 选项保存修改后的数据文件 **fictchol.psf**。

注意：

名义变量和有序变量在 PRELIS 中都被定义为有序变量。

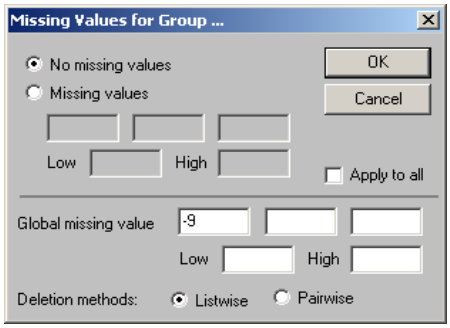
处理缺失值

对缺失值的处理是整理数据的重要环节。对其处理方法主要有两种：一是删除含缺失值的观测对象，或者填充缺失值。PRELIS 里，有两种方法删除含缺失值的对象：1、Listwise deletion （成列删除，即删除所有含缺失值的观测对象）；2、Pairwise deletion （成对删除，即计算两个变量的相关系数时，只使用两个变量都有数据的那些样品）。填补缺失值也有两个方法：1、匹配计算（impute by matching）；2、多元计算（multiple imputation）。这一小节，我们以图例说明如何定义缺失值。

定义整体缺失值（global missing value）并进行成列删除（listwise deletion）

通过以下步骤，我们将-9 定义为 **fitchol.psf** 中的整体缺失值。

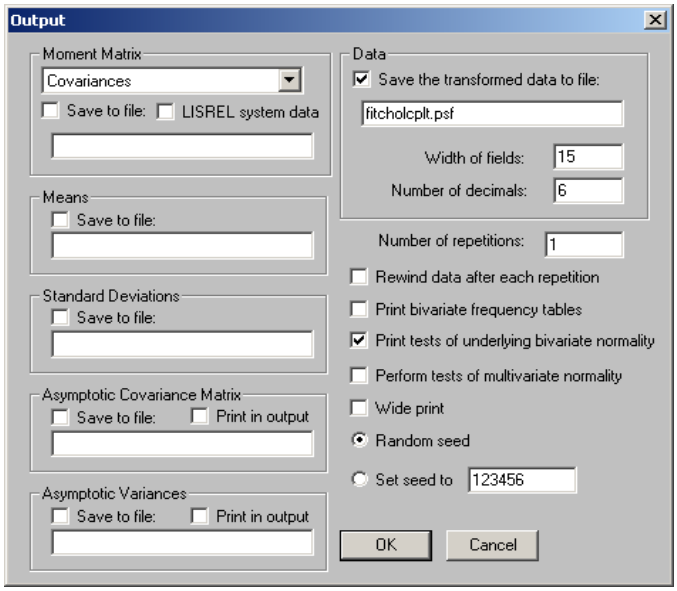
- 打开 PSF 窗口，点击 **Data** 菜单上的 **Define Variables** 选项。
- 在变量 列表中选择变量 Group 激活 **Define Variables** 对话框上的所有键。
- 点击 **Missing Values** 键打开 **Missing Values for Group…** 对话框。
- 如下所示在 **Global missing value** 对应的字符区键入 -9。
- 激活 **Deletion methods** 中的 **Listwise** 选项按钮。



- 点击 **OK** 键回到 **Define Variables** 对话框。
- 点击 **OK** 键 回到 PSF 窗口。
- 点击 **File** 菜单上的 **Save** 选项 保存 **fictchol.psf**。

我们可以将 listwise deletion 后没有任何缺失值的数据另存为一个新的数据文件。

- 点击 **Statistics** 菜单上的 **Output options** 选项。
- 钩上 **Data** 中 **Save the transformed data to file** 复选框。
- 如下图所示，在对应的字符区键入 **fitcholcplt.psf** 作为文件名。



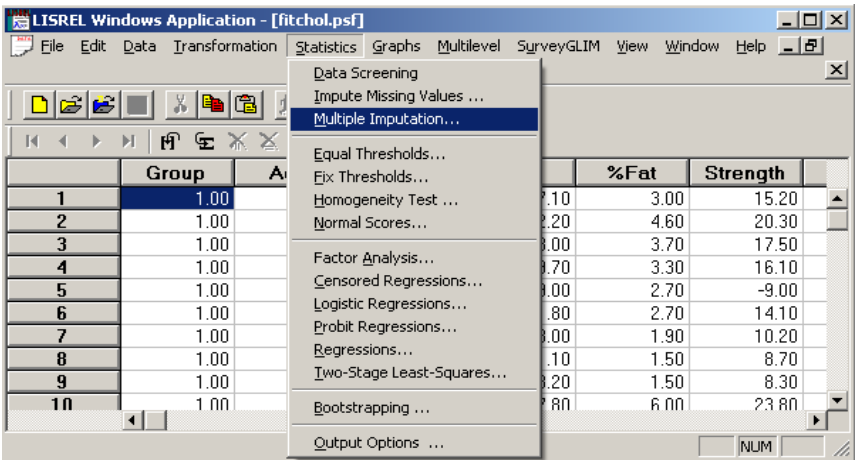
- 点击 **OK** 键 打开 **fitcholcplt.out** 窗口。

这样， **fitcholcplt.psf** 数据文件就生成了。它被存储在 **fitchol.psf** 所在的文件夹里。这个新的数据文件包含 57 个观测对象。

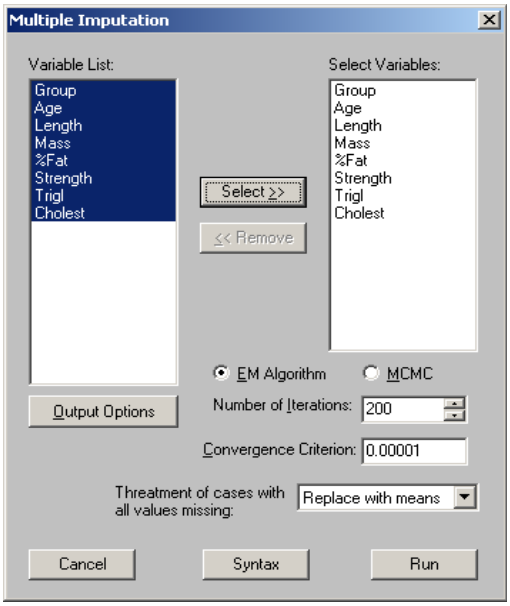
多元计算（multiple imputation）

有两种方法进行多元计算：Expected Maximization (EM 算法)和 Monte Carlo Markov Chain (MCMC 算法)。

- 点击 PSF 窗口中 **Statistics** 菜单上的 **Multiple Imputation** 选项激活 **Multiple Imputation** 对话框。



- 如下所示从变量列表中选择所有的变量并点击 **Select** 键。



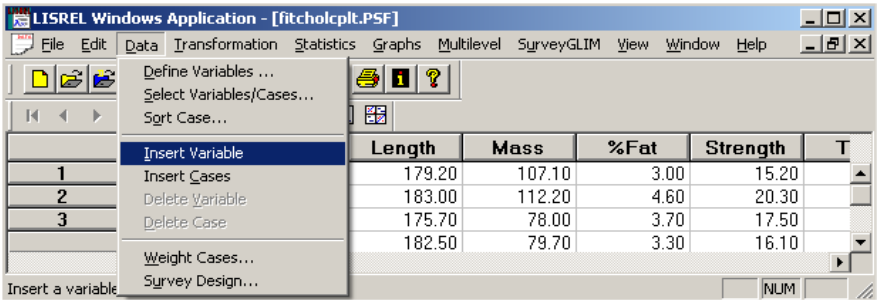
- 点击上述对话框中的 **Output Options** 打开 **Output** 对话框。
- 在 **Data** 部分中的 **Save the transformed data to file** 复选框打上钩。
- 在相应的字符区键入文件名 **fitcholimp.psf**。
- 点击 **OK** 键回到 **Multiple Imputation** 对话框。
- 点击 **Multiple Imputation** 对话框上的 **Run** 键运行 PRELIS 并生成输出文件 **fitchol.out**。

通过以上步骤，填补缺失值后的数据集 **fitcholimp.psf** 就生成了。这个文件存在 **fitchol.psf** 所在的文件夹中，包含 60 个观测对象，没有缺失值。

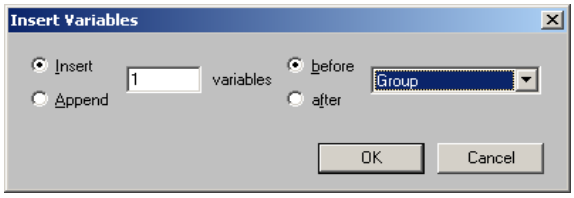
插入新变量

通过如下步骤，我们可以给 **fitcholcplt.psf** 数据集中插入一个新变量，并将其命名为 Totchol。

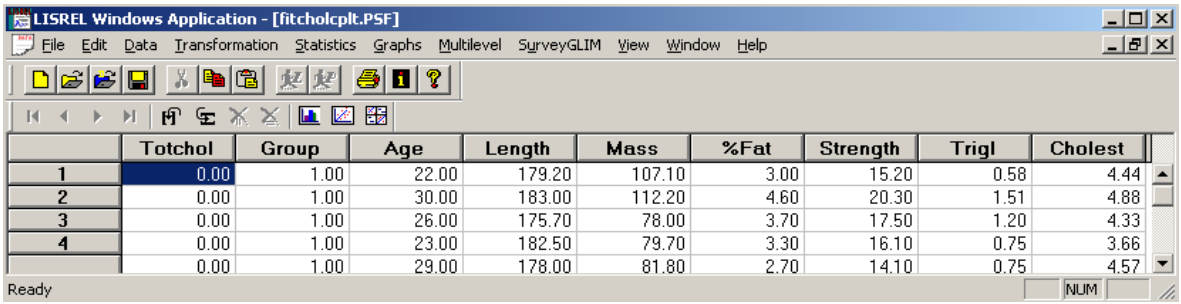
- 点击 **Windows** 菜单上的 **Close all** 选项关闭所有打开的窗口。
- 点击 **File** 菜单的 **Open** 选项。
- 在 **Files of type** 下拉菜单中选择 **PRELIS Data (*.psf)** 选项。
- 找到并选上 **fitcholcplt.psf** 文件，点击 **Open** 键 以打开 **fitcholcplt.psf** 的 PSF 窗口。
- 如下点击 **Data** 菜单上的 **Insert Variables** 选项



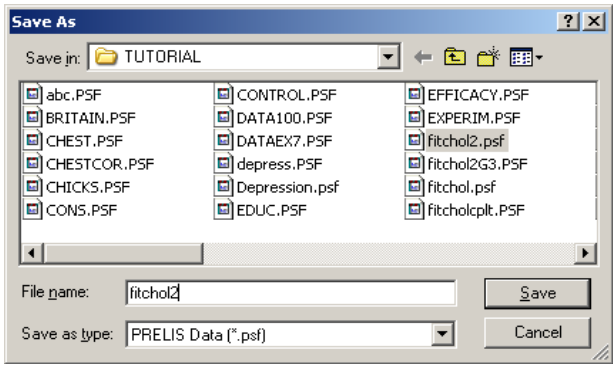
打开如下 **Insert Variables** 对话框。



- 点击 **OK** 键，一个新变量就被插入到数据文件中，其位置在 Group 前，变量名是 var9。
- 点击 **Data** 菜单 **Define Variables** 选项激活 **Define Variables** 对话框。
- 选中变量 var9。
- 点击 **Rename** 键。
- 键入 Totchol。
- 点击 **OK** 键回到 **Define Variables** 对话框。
- 点击 **Define Variables** 对话框中的 **OK** 键得到如下 PSF 窗口。



- 点击 **File** 菜单上的 **Save as** 选项。
- 如下在 **File name** 字符区键入 **fitchol2**。

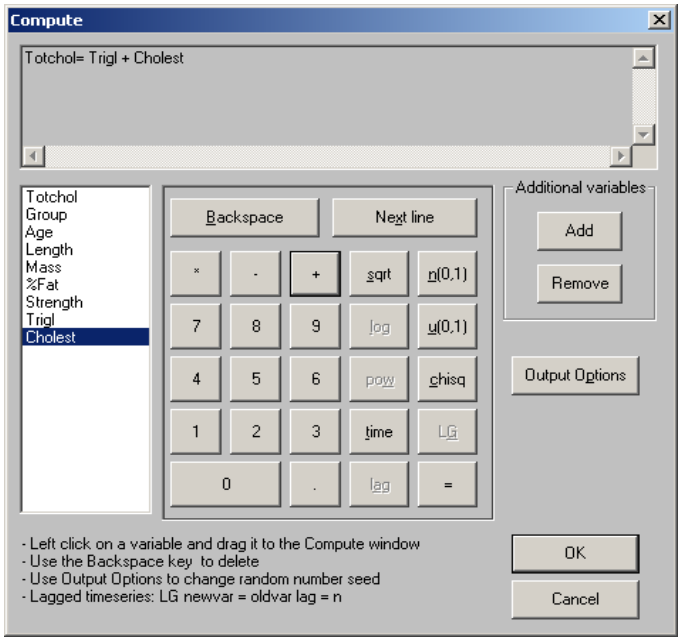


这样一个新变量 Totchol 就被插入到原有的数据集中并存储为 **fictchol2.psf** 。但是，这个变量的所有值都是零。

为新变量赋值

接下来，我们要使这个新变量 Totchol 代表 Trigl 和 Cholest 的和。

- 点击 **Transformation** 菜单上的 **Compute** 选项 打开 **Compute** 对话框。
- 选中并用鼠标将 Totchol 拖入 **Compute** 对话框中的灰色字符区。
- 点击 “**=**” 键。
- 选中并用鼠标将 Trigl 拖入 **Compute** 对话框中的灰色字符区。
- 点击 “**+**” 键。
- 选中并用鼠标将 Cholest 拖入灰色字符区得到如下 **Compute** 对话框。



- 点击 **OK** 键 可以看到如下 PSF 窗口。

fitchol2.psf

	Totchol	Group	Age	Length	Mass	%Fat	Strength	Trigl	Cholest
1	5.02	1.00	22.00	179.20	107.10	3.00	15.20	0.58	4.44
2	6.39	1.00	30.00	183.00	112.20	4.60	20.30	1.51	4.88
3	5.53	1.00	26.00	175.70	78.00	3.70	17.50	1.20	4.33
4	4.41	1.00	23.00	182.50	79.70	3.30	16.10	0.75	3.66
5	5.32	1.00	29.00	178.00	81.80	2.70	14.10	0.75	4.57
6	4.23	1.00	26.00	169.80	78.00	1.90	10.20	0.33	3.90
7	4.39	1.00	21.00	178.60	81.10	1.50	8.70	0.48	3.91
8	6.04	1.00	33.00	179.20	83.20	1.50	8.30	1.61	4.43
9	6.75	1.00	36.00	185.20	87.80	6.00	23.80	1.42	5.33
10	4.84	1.00	23.00	179.60	80.30	2.20	11.70	1.08	3.76

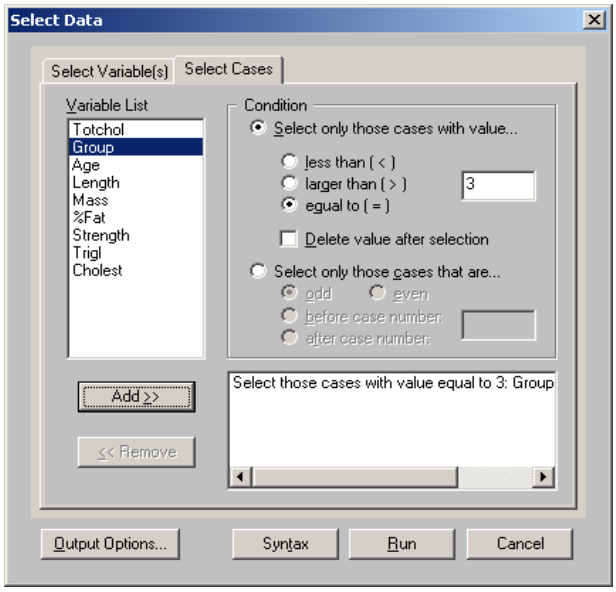
- 点击 **File** 菜单上的 **Save** 选项保存数据集 **fictchol2.psf**。

选择观测对象并产生一个子数据集

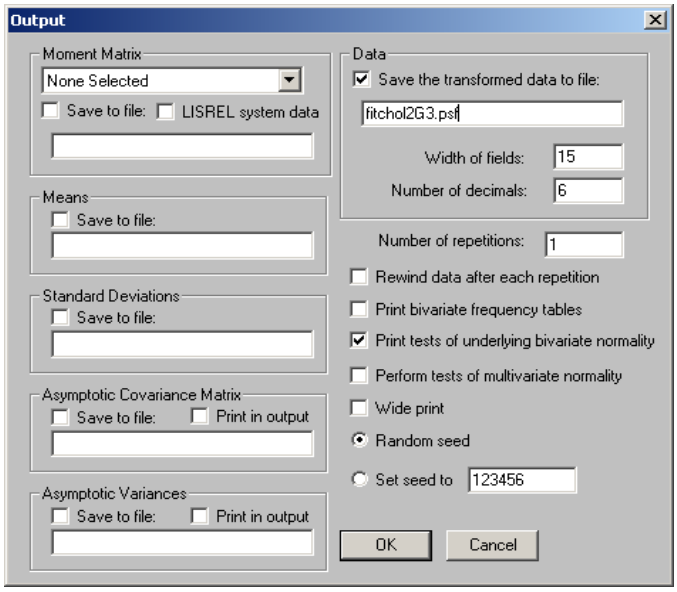
假如我们现在只想分析长跑运动员的数据。 我们可以将所有 Group 值为 3 的数据选出，并生成一个子数据集。

- 在 **fictchol2.psf** 窗口中，点击 **Data** 菜单上的 **Select Variables/Cases** 选项打开 **Select Data** 对话框。
- 点击 **Select Cases** 标签。
- 在 **Variable List** 列表里选中变量 Group。

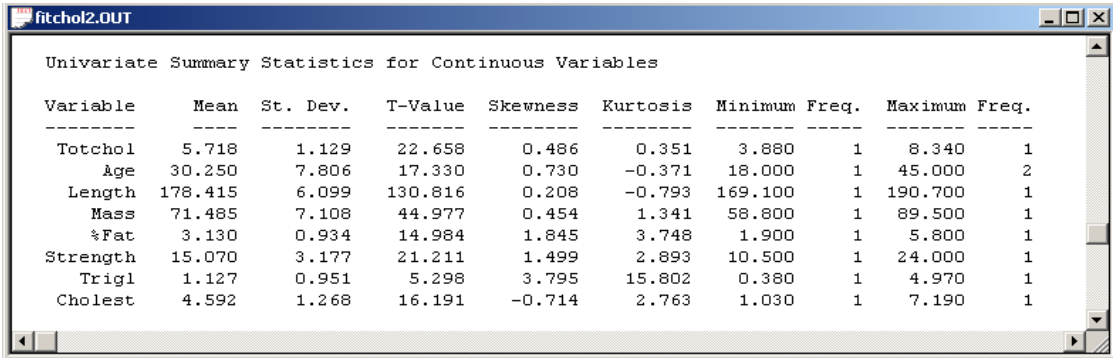
- 在 **Condition** 部分激活 **Select only those cases with vale** 选项按钮。
- 选中 **equal to (=)** 选项按钮。
- 在相应的字符区输入 3。
- 点击 **Add** 键得到如下 **Select Data** 对话框。



- 点击 **Output Variables** 打开 **Output** 对话框。
- 在 **Data** 区域中的 **Save the transformed data to file** 复选框打上勾。
- 如下图所示键入文件名 **fitchol2G3.psf**。



- 点击 **OK** 键 回到 **Select Data** 对话框。
- 点击 **Select Data** 对话框上的 **Run** 键生成结果文件 **fitchol2.out**。



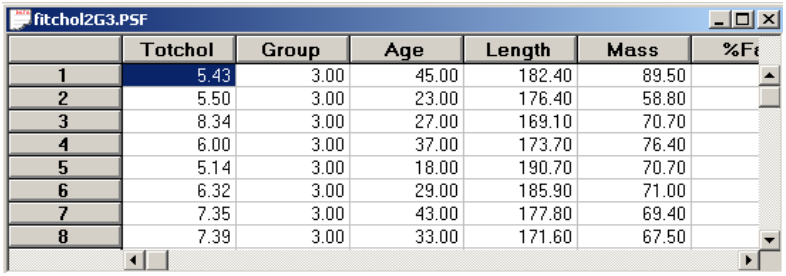
fitchol2.OUT

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
Totchol	5.718	1.129	22.658	0.486	0.351	3.880	1	8.340	1
Age	30.250	7.806	17.330	0.730	-0.371	18.000	1	45.000	2
Length	178.415	6.099	130.816	0.208	-0.793	169.100	1	190.700	1
Mass	71.485	7.108	44.977	0.454	1.341	58.800	1	89.500	1
%Fat	3.130	0.934	14.984	1.845	3.748	1.900	1	5.800	1
Strength	15.070	3.177	21.211	1.499	2.893	10.500	1	24.000	1
Trigl	1.127	0.951	5.298	3.795	15.802	0.380	1	4.970	1
Cholest	4.592	1.268	16.191	-0.714	2.763	1.030	1	7.190	1

通过如下步骤，可以看到新产生的数据集 **fitchol2G3.psf**。

- 首先，点击 **Windows** 菜单上 **Close all** 选项关闭所有打开的窗口。
- 点击 **File** 菜单 **Open** 选项。
- 点击 **PRELIS Data (*.psf)** 选项 from the **Files of type** 下拉菜单 box。
- 找到并选中 **fitchol2G3.psf**。
- 点击 **Open** 键 以打开如下 PSF 窗口。



fitchol2G3.PSF

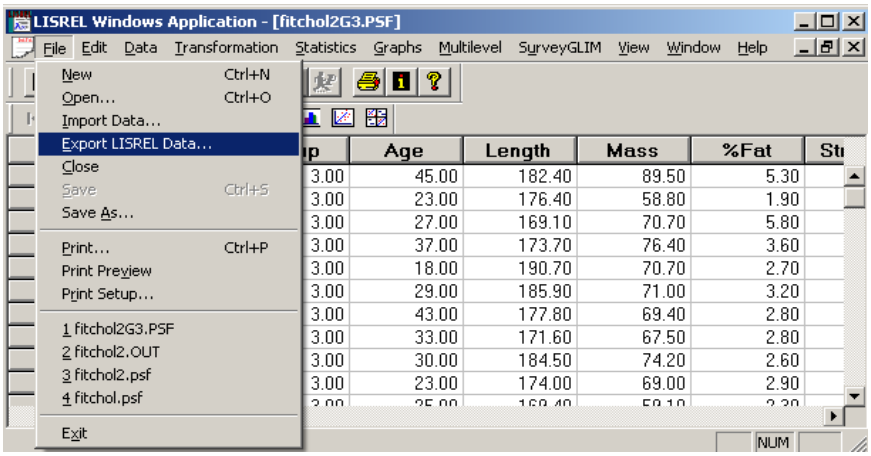
	Totchol	Group	Age	Length	Mass	%Fa
1	5.43	3.00	45.00	182.40	89.50	
2	5.50	3.00	23.00	176.40	58.80	
3	8.34	3.00	27.00	169.10	70.70	
4	6.00	3.00	37.00	173.70	76.40	
5	5.14	3.00	18.00	190.70	70.70	
6	6.32	3.00	29.00	185.90	71.00	
7	7.35	3.00	43.00	177.80	69.40	
8	7.39	3.00	33.00	171.60	67.50	

这个新的数据集中，只包括 20 个长跑运动员。

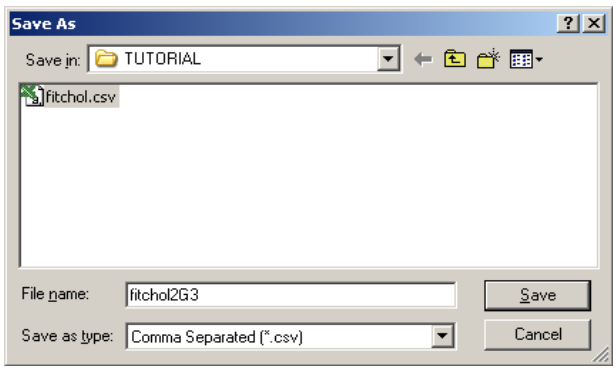
将 PSF 数据集输出为 Excel 可读的文件

我们用 **fitchol2G3.psf** 演示如何将 PSF 数据输出位 Excel 可读的文件。

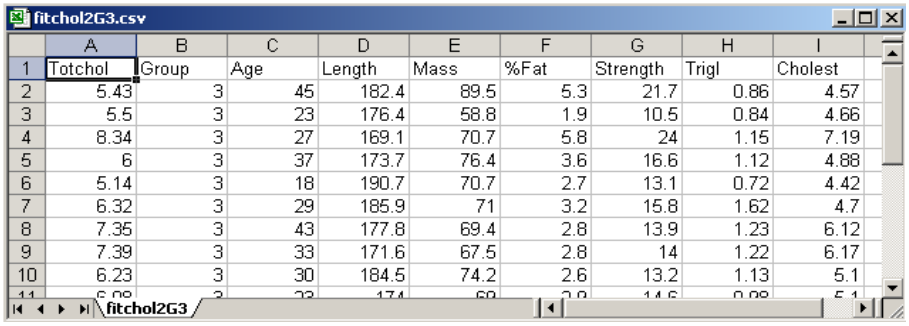
- 在 PSF 窗口中打开 **fitchol2G3.psf**。
- 点击 **File** 菜单上 **Export LISREL Data** 选项打开 **Save As** 对话框。



- 在 **Save as type** 下拉菜单中选择 **Comma Separated (*.csv)** 选项。
- 如下所示，在 **File name** 字符区键入 **fitchol2G3**。



- 点击 **Save** 键将数据集存储在 **TUTORIAL** 子文件夹中， 其文件名为 **fitchol2G3.csv**。
- 如下所示，在 **Excel** 表格窗口中打开这一文件。

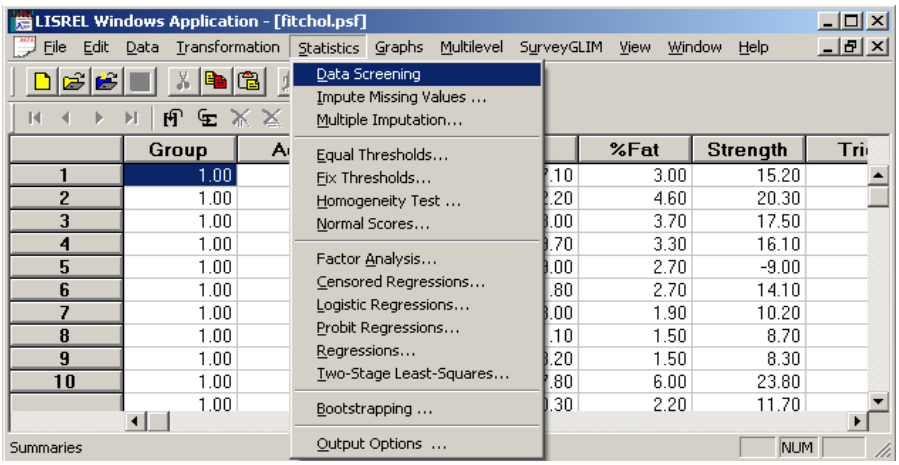


	A	B	C	D	E	F	G	H	I
1	Totchol	Group	Age	Length	Mass	%Fat	Strength	Trigl	Cholest
2	5.43	3	45	182.4	89.5	5.3	21.7	0.86	4.57
3	5.5	3	23	176.4	58.8	1.9	10.5	0.84	4.66
4	8.34	3	27	169.1	70.7	5.8	24	1.15	7.19
5	6	3	37	173.7	76.4	3.6	16.6	1.12	4.88
6	5.14	3	18	190.7	70.7	2.7	13.1	0.72	4.42
7	6.32	3	29	185.9	71	3.2	15.8	1.62	4.7
8	7.35	3	43	177.8	69.4	2.8	13.9	1.23	6.12
9	7.39	3	33	171.6	67.5	2.8	14	1.22	6.17
10	6.23	3	30	184.5	74.2	2.6	13.2	1.13	5.1

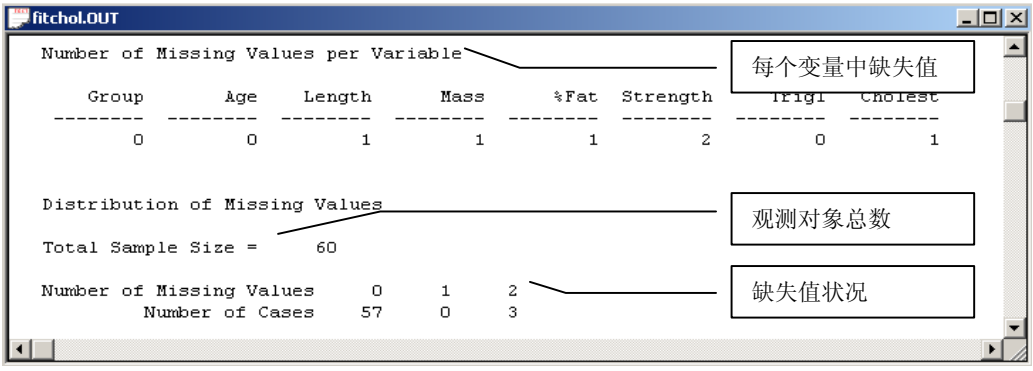
数据总览（data screening）

在进行数据分析之前，数据总览（data screening）是必不可少的一步。PRELIS 的数据总览功能提供了数据集的基本信息，比如：数据集中有多少变量，多少观测对象，缺失数据的状况等。同时，它还提供每个变量的简要信息。

- 打开数据集 **fictchol.psf**。
- 如下所示，点击 **Statistics** 菜单 **Data Screening** 选项。

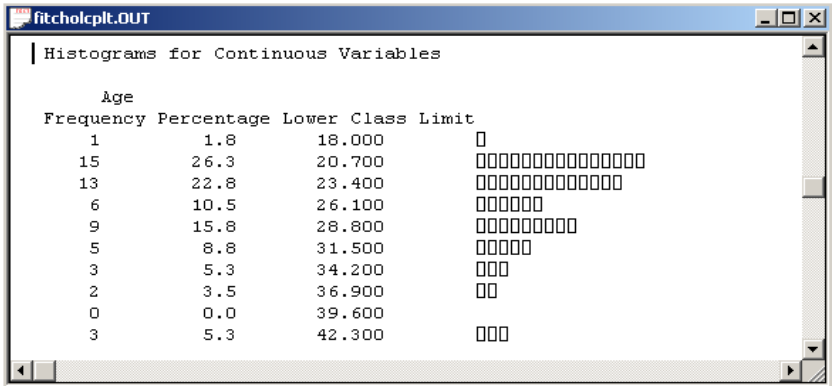


PRELIS 就会生成结果文件 **fitchol.out**。数据集的基本信息如下。



在上图的结果文件中可以看到：Group, Age 和 Trigl 中不含任何缺失值，Strength 中有 2 个缺失值，其他的变量中有一个缺失值。这个数据集中一共包含 60 个观测对象。不含任何缺失值的观测对象有 57 个。

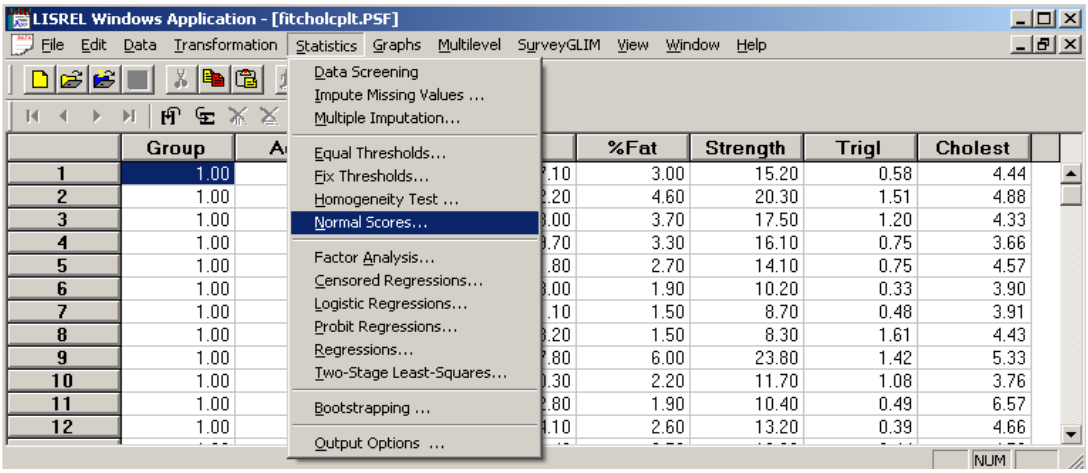
在这个结果文件中，我们还可以看到每个变量的基本信息。如下图所示，以 Age 为例，我们可以看到该变量的直方图。



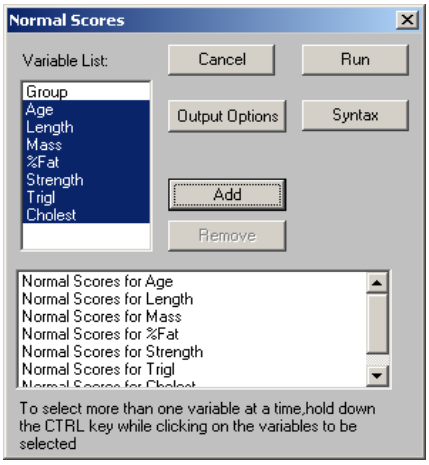
计算正态值（normal scores）

对于非正态分布的变量，我们可以用正态值使其正态化。 对于原点和度量单位没有实质意义的变量（比如考试成绩），使用其正态值是进行正态化的有效方法 。PRELIS 可以计算有序变量和连续变量的正态值。根据如下步骤，我们来计算 **fitcholcplt.psf** 中变量的正态值。

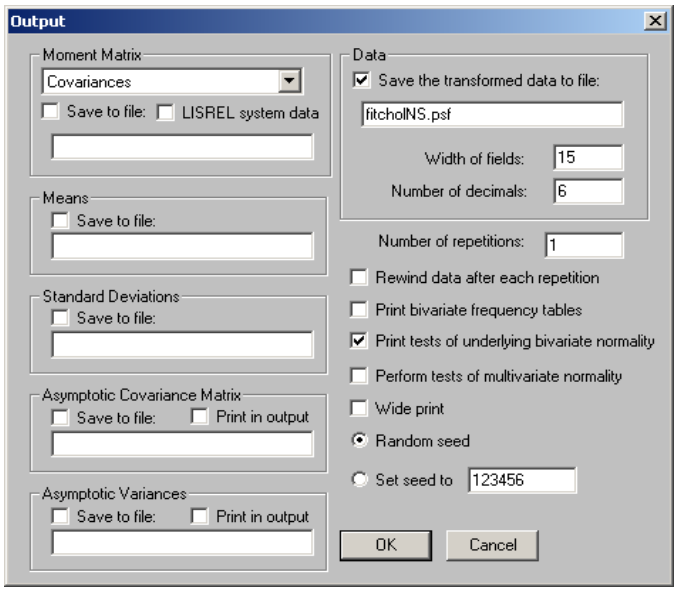
- 首先，点击 **Windows** 菜单 **Close all** 选项关闭所有打开的窗口。
- 点击 **File** 菜单的 **Open** 选项。
- 选择 **Files of type** 下拉菜单中的 **PRELIS Data (*.psf)** 选项。
- 找到并选择 **fitcholcplt.psf**。 （TUTORIAL 子文件夹中）
- 点击 **Open** 键 以打开 for **fitcholcplt.psf**。
- 如下所示，在 PSF 窗口中点击 **Statistics** 菜单 **Normal Scores** 选项 打开 **Normal Scores** 对话框。



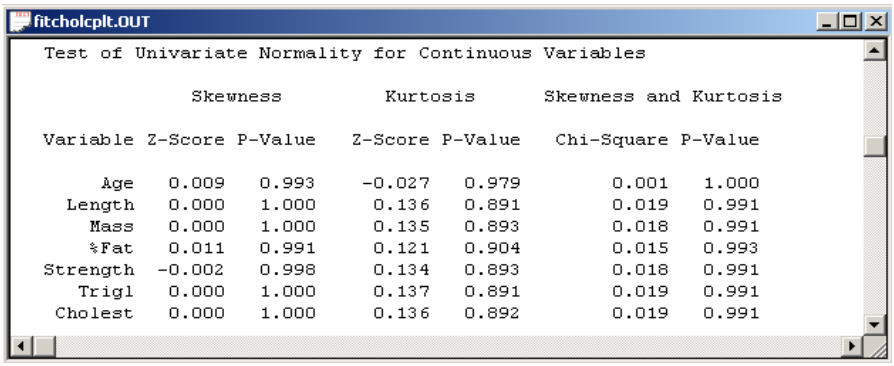
- 如下所示，在 **Variable List** 列表中选择所有的连续变量名并点击 **Add** 键。



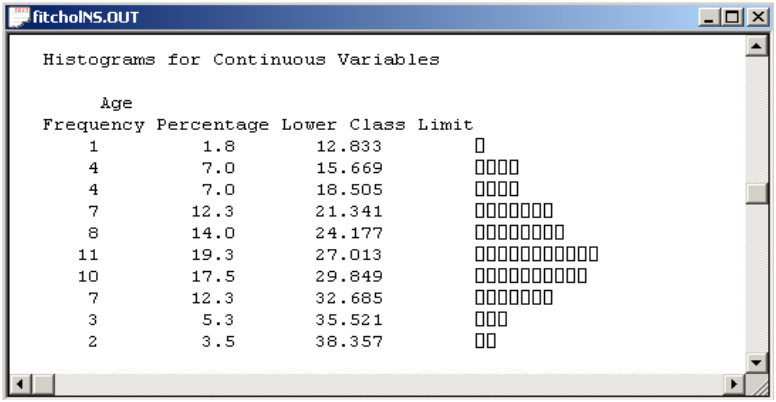
- 点击 **Output Options** 键。
- 如下所示，在 **Data** 部分 **Save the transformed data to file** 复选框中打上勾，并键入 **fitcholNS.psf** 作为文件名。



- 点击 **OK** 键 回到 **Normal Scores** 对话框。
- 点击 **Normal Scores** 对话框上的 **Run** 键打开如下 **fitcholcplt.out** 结果窗口。



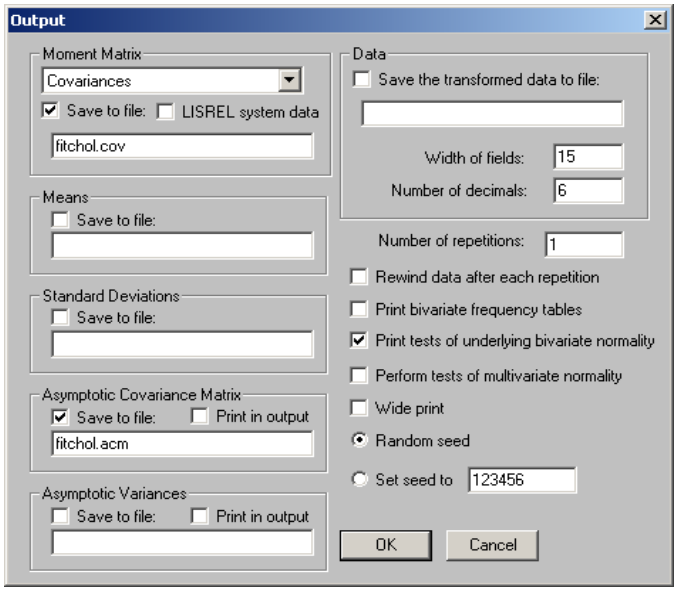
我们可以打开 **fitcholNS.psf** 数据集，并对它进行数据总览。Age 的直方图如下所示。



计算矩阵

在结构方程分析中常常会用到协方差矩阵，相关系数矩阵以及渐近协方差矩阵。我们以 **fitcholNS.psf** 数据集为例来示范如何计算协方差矩阵和渐近协方差矩阵。

- 点击 **Statistics** 菜单 **Output Options** 选项。
- 如下图所示，在 **Moment Matrix** 部分选择 **Covariances** 并在 **Save to file** 复选框中打上勾，输入 **fitchol.cov** 作为协方差矩阵的文件名。
- 如下图所示，在 **Asymptotic Covariance Matrix** 部分中的 **Save to file** 复选框中打上勾，输入 **fitchol.acm** 作为渐近协方差矩阵的文件名。



- 点击 **OK** 键运行 PRELIS 并打开 **fitcholcplt.out** 窗口。

通过以上步骤， **fitchol.cov** 就被计算并作为一个文本文件被存在 **fitcholcplt.psf** 所在的文件夹中。渐近协方差矩阵 **fitchol.acm** 作为一个二元(binary)文件也被存在同一个文件夹中。

例2：多元回归分析

关于数据

密歇根大学社会学学院 2001 年在 MTF 课题中，对 1608 所高中的在校学生进行了饮酒和毒品使用的调查。我们应用此项调查的部分数据生成了数据集 **select.psf**。如下图所示，是这个数据集的前 10 个样本。

select.PSF

	school	region	alclifs	alcanns	alc30ds	xmjlifs	xmj12mos	xmj30ds	tick12mo	acci12mo	newwgt	wt
1	5.00	1.00	7.00	7.00	5.00	7.00	7.00	6.00	-999999.00	2.00	9.52	1.15
2	5.00	1.00	6.00	3.00	2.00	7.00	6.00	3.00	0.00	0.00	9.52	1.15
3	5.00	1.00	3.00	2.00	1.00	3.00	2.00	2.00	1.00	1.00	9.52	1.15
4	5.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	9.52	1.15
5	5.00	1.00	6.00	3.00	1.00	1.00	1.00	1.00	1.00	2.00	9.52	1.15
6	5.00	1.00	5.00	4.00	4.00	6.00	1.00	1.00	0.00	0.00	9.52	1.15
7	5.00	1.00	6.00	5.00	3.00	5.00	4.00	2.00	2.00	0.00	9.52	1.15
8	5.00	1.00	3.00	3.00	1.00	1.00	1.00	1.00	0.00	0.00	9.52	1.15
9	5.00	1.00	7.00	3.00	1.00	7.00	6.00	3.00	0.00	1.00	9.52	1.15
10	5.00	1.00	5.00	4.00	3.00	2.00	1.00	1.00	0.00	1.00	9.52	1.15

注意 -999999.00 是系统默认的缺失值。这个数据集中的部分变量包括：

- alclifs 是对问题 “你迄今为止共饮酒多少次？ ”的量化回答。

- alc12mos 是对问题 “你在过去的 12 个月中共饮酒多少次？”的量化回答。
- alc30ds 是对问题 “你在过去的 30 天中共饮酒多少次？”的量化回答。
- xmjlifs 是对问题 “你迄今为止共服用过多少次毒品（大麻）？ ”的量化回答。
- xmj12mos 是对问题 “你过去的 12 个月中共服用过多少次毒品（大麻）？”的量化回答。
- xmj30ds 是对问题 “你在过去的 30 天中共服用过多少次毒品（大麻）？”的量化回答。
- tick12mo 是对问题 “你过去的 12 个月中共吃过多少张驾驶违规的罚单？”的量化回答。
- acci12mo 是对问题 “你过去的 12 个月中共发生过几次交通事故？”的量化回答。

关于这项课题的详细信息，请参考如下网址。
http://www.icpsr.umich.edu/cgi/archive.prl?study=3088&path=SAMHDA.

多元线性回归模型

多元线性回归常用于研究一组自变量 $x_1,\cdots,x_p$ 和一个因变量 y 之间的关系，其数学表达式可通常写作

$$y=\beta_0+\beta_1x_1+\cdots+\beta_px_p+\varepsilon$$

β_0 代表截距，$\beta_1,\cdots,\beta_p$ 是回归系数，ε 代表误差。这个例子中，多元线性回归模型可以写作

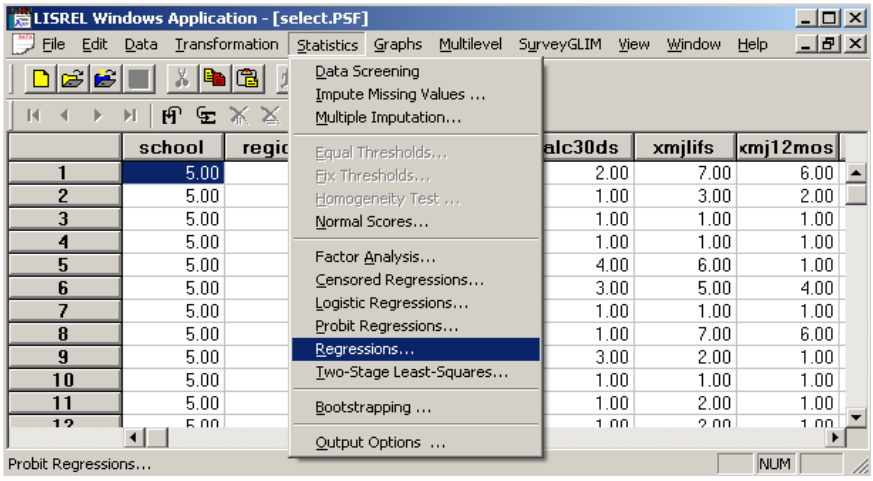
$$\text{acci12mo}=\beta_0+\beta_1\text{alclifs}+\beta_2\text{alcanns}+\beta_3\text{alc30ds}+\beta_4\text{xmjlifs}+\beta_5\text{xmj12mos}+\beta_6\text{xmj30ds}+\varepsilon$$

β_0 则代表不饮酒不吸烟的观测对象在过去的 12 个月中平均得到罚单的数量，$\beta_1,\cdots,\beta_6$ 是未知的回归系数，ε 代表误差.

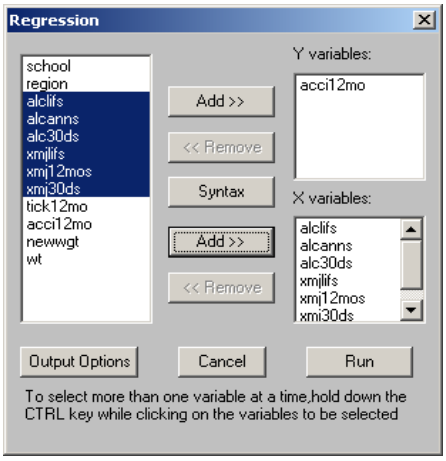
多元线性回归分析

通过如下步骤，我们对数据 **select.psf** 构建多元线性模型。

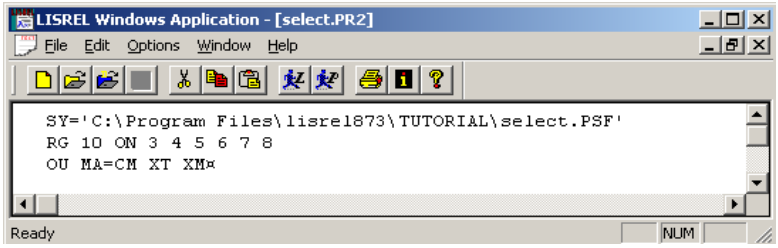
- 首先在 PSF 窗口打开 **TUTORIAL** 子文件夹中的 **select.psf** 数据集。
- 如下图所示，选择 **Statistics** 菜单上的 **Regressions** 选项打开 **Regression** 对话框。



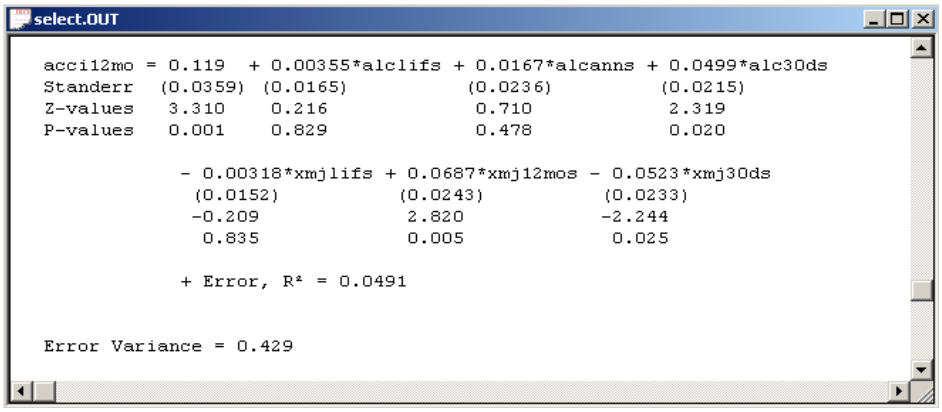
- 从变量列表中选择 acci12mo 并点击靠上的 **Add** 键，把这个变量加入 **Y variables** 列表中。
- 再从变量列表中选择 alclifs, alcanns, alc30ds, xmjlifs, xmj12mos, 和 xmj30ds，点击靠下的 **Add** 键得到如下对话框。



- 可以点击 **Syntax** 键，生成如下 **select.pr2** 程序文件。然后点击键来运行该文件。



- 也可以直接点击 **Regression** 对话框上的 **Run** 键得到如下结果文件 **fitchol.out**。



例3：用美国经济数据构建二阶最小二乘模型

二阶最小二乘(Two-stage Least Squares TSLS) 模型常常可以提供信息来判断结构模型是否合理。

关于数据

Klein 的模型 I (Klein 1950) 是一个经典的经济计量模型。这个模型包含 8 个等式，使用美国在两次世界大战间的经济数据。这是一个动态模型，时间在这个模型中有重要的意义。

数据文件 **klein.psf** 如下图所示，存储在 **TUTORIAL** 子文件夹中。我们用这一数据集构建一个简单的二阶最小二乘模型。

KLEIN.PSF

	Ct	Pt-1	Wt*	It	Kt-1	Et-1	Wt**	Tt	At	Pt	Kt	Et	Wt	Yt	Gt
1	41.9	12.7	25.5	-0.2	182.8	44.9	2.7	7.7	-10.0	12.4	182.6	45.6	28.2	40.6	6.6
2	45.0	12.4	29.3	1.9	182.6	45.6	2.9	3.9	-9.0	16.9	184.5	50.1	32.2	49.1	6.1
3	49.2	16.9	34.1	5.2	184.5	50.1	2.9	4.7	-8.0	18.4	189.7	57.2	37.0	55.4	5.7
4	50.6	18.4	33.9	3.0	189.7	57.2	3.1	3.8	-7.0	19.4	192.7	57.1	37.0	56.4	6.6
5	52.6	19.4	35.4	5.1	192.7	57.1	3.2	5.5	-6.0	20.1	197.8	61.0	38.6	58.7	6.5
6	55.1	20.1	37.4	5.6	197.8	61.0	3.3	7.0	-5.0	19.6	203.4	64.0	40.7	60.3	6.6
7	56.2	19.6	37.9	4.2	203.4	64.0	3.6	6.7	-4.0	19.8	207.6	64.4	41.5	61.3	7.6
8	57.3	19.8	39.2	3.0	207.6	64.4	3.7	4.2	-3.0	21.1	210.6	64.5	42.9	64.0	7.9
9	57.8	21.1	41.3	5.1	210.6	64.5	4.0	4.0	-2.0	21.7	215.7	67.0	45.3	67.0	8.1
10	55.0	21.7	37.9	1.0	215.7	67.0	4.2	7.7	-1.0	15.6	216.7	61.2	42.1	57.7	9.4

这个数据文件中包含以下 15 个变量。

- Ct 是 t 年的总消费
- Pt-1 是 前一（t-1）年的总利润
- Wt* 代表 t 年的个人收入总和
- It 是 t 年的净投资总和
- Kt-1 是前一（t-1）年的资本总和
- Et-1 是前一（t-1）年的私有工业总产值
- Wt** 是 t 年的政府工资支出
- Tt 是 t 年的税收
- At 是从 1931 算起的时间, 即 t-1931
- Pt 是 t 年的总利润
- Kt 是 t 年年底的资本总和
- Et 是 t 年私有工业总产值
- Wt 是 t 年的工资支出
- Yt 是 t 年的个人收入总和
- Gt 是 t 年的政府非工资支出

除 At 外，其他变量均以 1934 的美元为基准，十亿为单位。

二阶最小二乘数学模型

二阶最小二乘(Two-stage least squares TSLS) 在经济计量学中常常会用到。其数学表达式为

$$\mathbf{y} = \mathbf{B}\mathbf{y} + \mathbf{\Gamma}\mathbf{x} + \mathbf{u}$$

$\mathbf{y} = (y_1, y_2, ..., y_p)'$ 是一组内生变量。 $\mathbf{x} = (x_1, x_2, ..., x_q)'$ 是一组外生变量。 $\mathbf{u} = (u_1, u_2, ..., u_p)'$ 代表误差，它与 $\mathbf{x}$ 相互独立。 $\mathbf{B}$ 和 $\mathbf{\Gamma}$ 是系数矩阵。

这个模型的重要特点在于不是每个等式都包括所有的 *y*-变量和所有的 *x*-变量。对模型的每一个等式，可识别的必要条件是对每一个等式右侧的 *y*-变量，至少要有一个 *x*-变量不被包含在这个等式中。可识别的充分条件是所谓的秩条件（rank condition） 但是这一条件在实际应用中往往很难满足。更多的信息可以参考 Goldberger (1964, pp. 313-318)。

我们将要构建的模型可以表达如下：

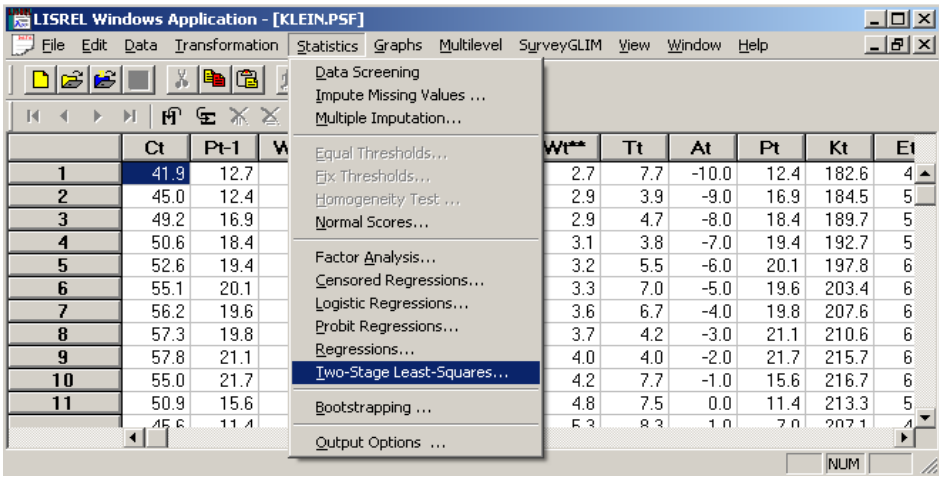
$$C_t = \beta_0 + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 W_t + \varepsilon_t$$

β_0 代表 t 年的平均总消费，β_1，β_2 和 β_3 未知的系数，ε_t 代表误差。

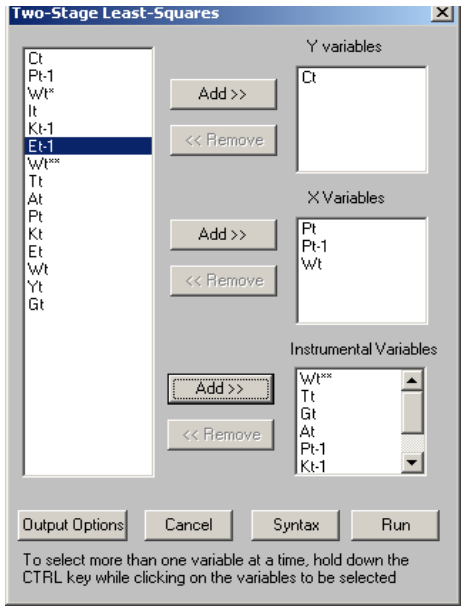
二阶最小二乘分析

通过如下步骤，我们对数据 **klein.psf** 构建二阶最小二乘模型。

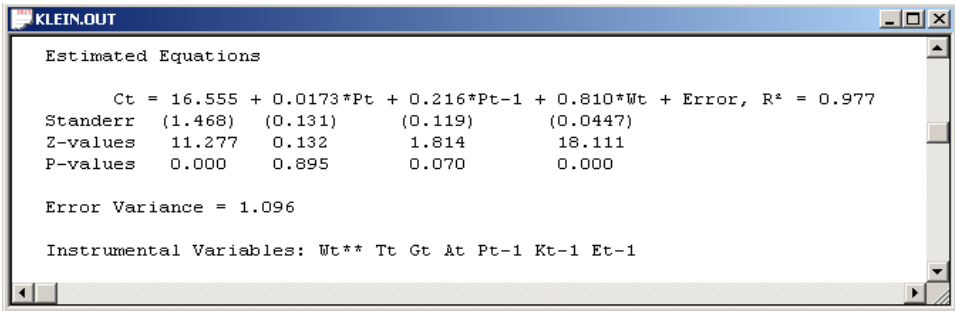
- 首先在 PSF 窗口打开 **TUTORIAL** 子文件夹中的 **select.psf** 数据集。
- 如下图所示，选择 **Statistics** 菜单上的 **Two-Stage Least-Squares** 选项打开 **Two-Stage Least-Squares** 对话框。



- 从变量列表中选择 Ct 并点击靠上的 **Add** 键，把这个变量加入 **Y variables** 列表中。
- 再从变量列表中选择 Pt, Pt-1 和 Wt 并点击中间的 **Add** 键，把这个变量加入 **X variables** 列表中。
- 最后从变量列表中选择 Wt**, Tt, Gt, At, Pt-1, Kt-1 和 Et-1acci12mo 并点击靠下的 **Add** 键，把这个变量加入 **Instrumental variables** 列表中。



- 点击对话框上的 **Run** 键得到如下结果文件 **fitchol.out**。



例4 ： 以心理学数据为例示范探索性因子分析

探索性因子分析常被用于构建结构模型之前对数据进行因子分析。探索性分析可以发现变量之间的关系，提供建立结构、模型的线索，假设检验的设定。PRELIS 提供三种系数估计：未旋转的 ML 估计， VARIMAX 旋转的结果和 PROMAX 旋转的结果。因子的个数可以由用户决定，也可以由程序自选。

关于数据

我们用 Holzinger 和 Swineford (1939)的经典数据集来示范探索性因子分析。在他们的数据集中，收集到芝加哥一所白人中学里 145 个 7 年级 8 年级的学生的问卷测试。我们取其中的 9 个项目的测试成绩生成 **npv.psf** 数据集。 前 10 个被试的观测值如下图所示。

NPV.PSF

	VISPERC	CUBES	LOZENGES	PARCOMP	SENCOMP	WORDMEAN	ADDITION	COUNTDOT	SCCAPS
1	23.00	19.00	4.00	10.00	17.00	10.00	69.00	82.00	156.00
2	33.00	22.00	17.00	8.00	17.00	10.00	65.00	98.00	195.00
3	34.00	24.00	22.00	11.00	19.00	19.00	50.00	86.00	228.00
4	29.00	23.00	9.00	9.00	19.00	11.00	114.00	103.00	144.00
5	16.00	25.00	10.00	8.00	25.00	24.00	112.00	122.00	160.00
6	30.00	25.00	20.00	10.00	23.00	18.00	94.00	113.00	201.00
7	36.00	33.00	36.00	17.00	25.00	41.00	129.00	139.00	333.00
8	28.00	25.00	9.00	10.00	18.00	11.00	96.00	95.00	174.00
9	30.00	25.00	11.00	11.00	21.00	8.00	103.00	114.00	197.00
10	20.00	25.00	6.00	9.00	21.00	16.00	89.00	101.00	178.00
	27.00	26.00	6.00	10.00	16.00	13.00	88.00	107.00	137.00

- VISPERC 代表视觉敏锐考试的成绩（视觉考试之一）。
- CUBES 代表立方体考试的成绩（视觉考试之一）。
- LOZENGES 是菱形考试的成绩（视觉考试之一）。
- PARCOMP 是段落完整性考试的成绩（语文考试之一）。
- SENCOMP 是句子完整性考试的成绩（语文考试之一）。
- WORDMEAN 是单词意思考试的成绩（语文考试之一）。
- ADDITION 是加法考试的成绩（速度考试之一）。
- COUNTDOT 是数数考试的成绩（速度考试之一）。
- SCCAPS 辨别字母考试的成绩（速度考试之一）。

以上 9 项考试成绩均应按连续性变量处理。

探索性因子分析的数学模型

探索性因子分析的目的是找到一组外源变量 $x_1,\cdots,x_q$ 中隐含的潜变量 $\xi_1,\cdots,\xi_n$。这里，潜变量的数目要小于外源变量，即 $n<q$。这些潜变量（因子）可以为构建结构模型提供参考。其数学模型可以写为：

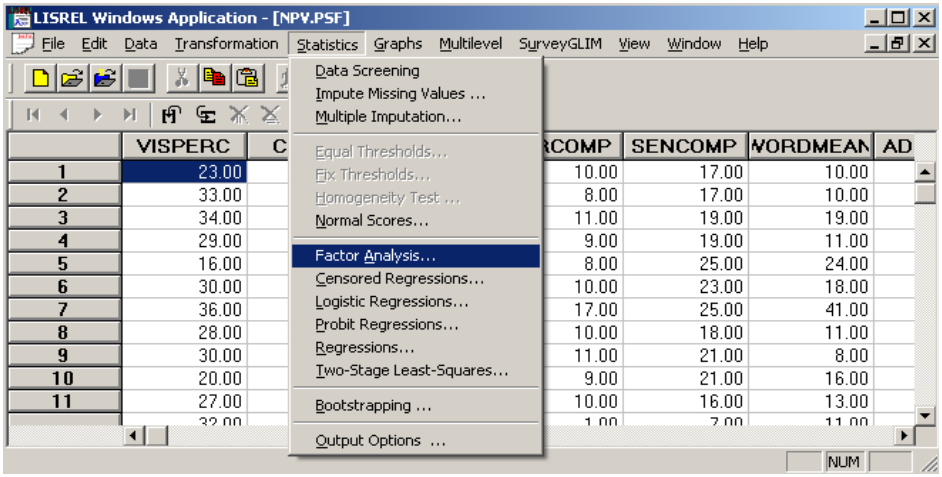
$$\mathbf{x}=\mathbf{\Lambda}\xi+\delta$$

这里我们假设 $\mathrm{E}(\xi)=\mathbf{0}$， $\mathrm{E}(\delta)=\mathbf{0}$，$\delta$ 和 ξ 相互独立。

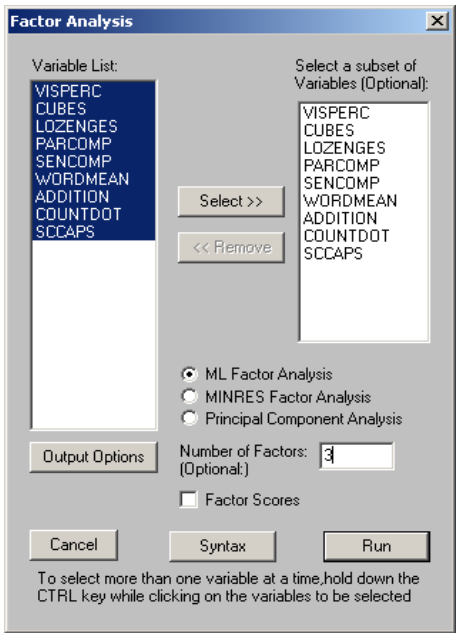
并不是所有类型的变量都适合探索性因子分析。比如性别，物质状态等描述特征的变量就不适合这一分析。在我们的例子中，9 项考试成绩将都被用于探索性分析。

探索性因子分析

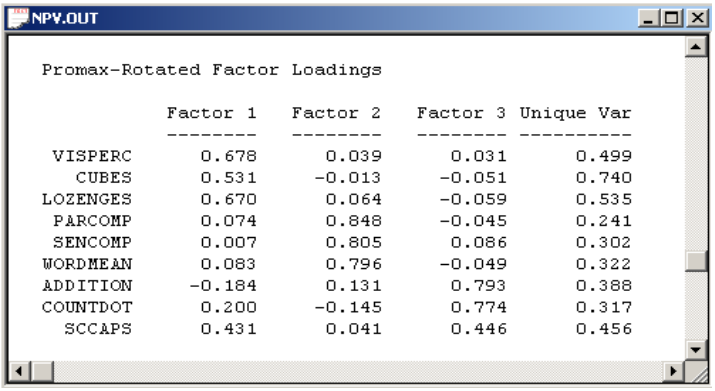
- 点击 **File** 菜单上的 **Open** 选项。
- 选择 **Files of type** 下拉菜单中的 **PRELIS Data (*.psf)** 选项。
- 选中 **TUTORIAL** 文件夹中的 **npv.psf**。
- 点击 **Open** 键在 PSF 窗口中打开 **npv.psf**。
- 如下图所示，选择 **Statistics** 菜单上的 **Factor Analysis** 选项。



- 选择 **Variable List:**列表中的所有变量。
- 点击 **Select** 键将选中的变量移到 **Select a subset of Variables** 列表中。
- 激活 **ML Factor Analysis** 并在 **Number of factors** 中键入 3。



- 点击 **Run** 键，我们就可以看到如下的 **npv.out** 结果文件。


NPY.OUT
Promax-Rotated Factor Loadings
Factor 1 Factor 2 Factor 3 Unique Var
VISPERC 0.678 0.039 0.031 0.499
CUBES 0.531 -0.013 -0.051 0.740
LOZENGES 0.670 0.064 -0.059 0.535
PARCOMP 0.074 0.848 -0.045 0.241
SENCOMP 0.007 0.805 0.086 0.302
WORDMEAN 0.083 0.796 -0.049 0.322
ADDITION -0.184 0.131 0.793 0.388
COUNTDOT 0.200 -0.145 0.774 0.317
SCCAPS 0.431 0.041 0.446 0.456