PASW Statistics CDB & R

A demo with the *polycor R* package

João Maroco, Ph. D.
Instituto Superior de Psicologia Aplicada
jpmaroco@ispa.pt

ASSESS Meeting. York, October 8-9, 2009

SUMMARY

1. How to measure association between ordinal variables: Pearson, Spearman or Polychoric?
2. The Polychoric correlation: theoretical background
3. Where can we get Polychoric correlations: The R polycor package
4. How about PASW? The PASW Statistics R integration package
5. Getting the polycor into PASW Statistics menus: The PASW Custom Dialog Builder

1.

How to measure association between ordinal variables:

Pearson, Spearman or Polychoric?

Ordinal variables

are quite frequent in the social, biomedical and engineering sciences:

- Likert scales, e.g. Agreement with social policies: [Completely Disagree; Disagree a little; Neither Agree or Disagree; Agree; Completely agree]
- Pain severity [It hurts a lot; It hurts; It hurts a little; Doesn't hurt]
- Concrete cracking [No cracks; a very few cracks; some cracks; lots of cracks]
- Mineral Hardiness [1 – Talc; 2 – Gypsum, 3 – Calcite; … ;10 – Diamond]

Ordinal variables (following Stevens, 1946):

- Reflect qualitative appreciations
- Have ordering properties
- But quantification of categories do not make sense
- Numbers, generally, assigned to categories only reflect ranks and not quantities

How do we measure correlations between ordinal variables?

1. **Pearson Product-moment correlation**: frequently used when categories are coded numerically with a ranking nature and linear relationship assumed:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{Cov(X,Y)}{S_X S_Y}$$

   Controversial:
   - Requires calculations of Means and SD
   - Quite frequent in Psychometrics. Stevens criticized its use in this field:
     "As a matter of fact, most of the scales used widely and effectively by psychologists are ordinal scales. In the strictest propriety the ordinary statistics involving means and standard deviations ought not to be used with these scales, for these statistics imply a knowledge of something more than the relative rank order of data" (Stevens, 1946, p. 679).
   - 'Tolerated' when the number of categories are large (5, 7 or even 10)

2. **Spearman Correlation**: Proposed by British psychologist Charles Spearman.
   Assesses how well an arbitrary monotonic function (ranking) could describe the relationship between variables, without making any assumptions about the nature of the relationship:

If no tied ranks: $$R_S = 1 - \frac{6\sum_{i=1}^{n}(R_{X_i} - R_{Y_i})^2}{n(n^2-1)}$$

If tied ranks: $$R_S = \frac{\sum_{i=1}^{n}(R_{X_i} - \bar{R}_X)(R_{Y_i} - \bar{R}_Y)}{\sqrt{\sum_{i=1}^{n}(R_{X_i} - \bar{R}_X)^2 \sum_{i=1}^{n}(R_{Y_i} - \bar{R}_Y)^2}}$$

- It's just Pearson correlation on ranks.
- Frequently used for Descriptive Statistics
- But not for multivariate correlacional techniques like Factor Analysis (which, by the way, was first proposed by Spearman, 1904).

3. Polychoric correlation

Psychometricians theorize that most psychometric tests produce interval scale measures of cognitive abilities.

Thus, ordinal items in psychometric scales are just a ‘practical’ operationalization of real continuous (latent) variables which can only be assed by means of ordinal scaled (manifest) categories.

To assess the ‘true’ construct and their associations, we should estimate the correlation between the latent continuous variables operationalized by the ordinal items

Polychoric correlations does that. Early work done by Pearson, who acknowledge the facility of measuring variables in ordered categories

[Polyserial: one variable ordinal and one variable quantitative]

[Tethracoric: same as polychoric but for dichotomous variables]

Which correlation coefficient should one use?

Babakus, Ferguson, & Jöreskog (1987) in the context of Confirmatory Factor Analysis:

> "*The polychoric correlation [as compared to Product-moment, Spearman's rho, and Kendall's tau] procedure was found to provide the most accurate estimates of pairwise correlations and factor loadings*"

2.

The Polychoric Correlation:

Theoretical Background

What is a Polychoric correlation?

Estimate of the association between two latent normally distributed continuous latent variables, from which the two ordinal observed variables are manifestations.

Example:

Association between

'X_1 - Satisfaction with service' (from '1- Very dissatisfied' to '5 – Completely satisfied')

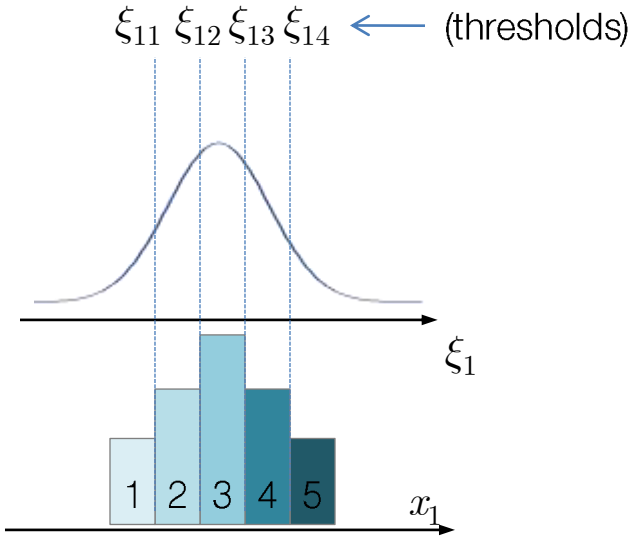
and

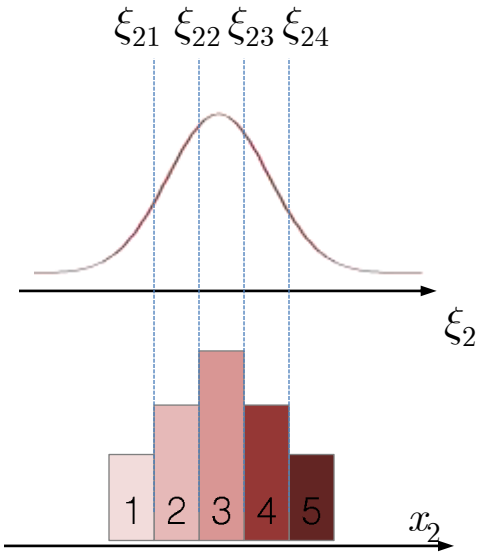
'X_2-Recommend service to a friend' (from '1- Definitively not' to '5- Definitively yes')

Lets assume:

1. 'Satisfaction' is a (approximately) normally distributed latent continuous variable, for which X_1 is a categorical ordered operationalization
2. 'Recommendation' is (approximately) normally distributed latent continuous variable, for which X_2 is a categorical ordered operationalization

operationalized as:

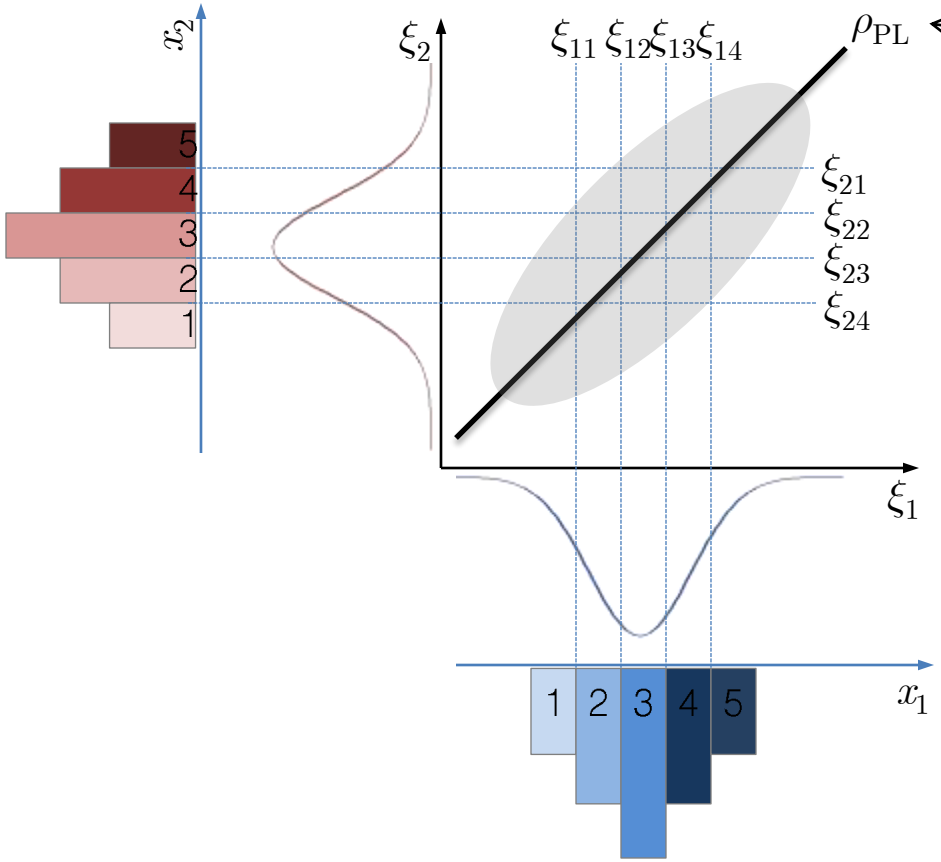




$$X_1 = \begin{cases} \text{'1 – Completely dissatisfied'} \ if \ \xi_1 \leq \xi_{11} \\ \text{'2 – Dissatisfied'} \ if \ \xi_{11} < \xi_1 \leq \xi_{12} \\ \text{'3 – Indiferent' se} \ \xi_{12} < \xi_1 \leq \xi_{13} \\ \text{'4 – Satisfied'} \ if \ \xi_{13} < \xi_1 \leq \xi_{14} \\ \text{'5 – Completely Satisfied'} \ if \ \xi_1 > \xi_{14} \end{cases}$$

$$X_2 = \begin{cases} \text{'1 – Definitively not'} \ if \ \xi_2 \leq \xi_{21} \\ \text{'2 – Probably not'} \ if \ \xi_{21} < \xi_2 \leq \xi_{22} \\ \text{'3 – Maybe' se} \ \xi_{22} < \xi_2 \leq \xi_{23} \\ \text{'4 – Probably yes'} \ if \ \xi_{23} < \xi_2 \leq \xi_{24} \\ \text{'5 – Definitively yes'} \ if \ \xi_2 > \xi_{24} \end{cases}$$

Combining the two variables:



The **Polychoric Correlation** (ρ_{PL}) is the product-moment correlation between the two latent normally distributed variables

How do we estimate Polychoric correlations?

0. Polychoric mathematical series decomposition (not in use anymore)

1. Maximum Likelihood:
   a. Estimate combined probability for observations x_{1i} and x_{2i} from standard normally distributed ξ_1 and ξ_2 as

$$P_{ij} = \int_{\xi_{1i-1}}^{\xi_{1i}} \int_{\xi_{2i-1}}^{\xi_{2i}} \phi(\xi_1, \xi_2; \rho) d\xi_2 d\xi_1 \quad \text{where} \quad \phi(\xi_1, \xi_2; \rho) = \frac{1}{2\pi\sqrt{(1-\rho^2)}} \times e^{\left(\frac{-1}{2(1-\rho^2)}(\xi_1^2 - 2\rho\xi_1\xi_2 + \xi_2^2\right)}$$

   is the standard normal bivariate density of ξ_1 and ξ_2

   b. Obtain Maximum likelihood estimate of ρ_{PC}

$$L = k\prod_{i=1}^{r}\prod_{j=1}^{s} P_{ij}^{n_{ij}}$$

where n_{ij} – number of observations of x_{1i} e x_{2j};
r, s – number of ordered categories of ξ_1 and ξ_2.

Differentiate the $Ln(L)$ with respect to all model parameters (ρ, ξ_{11}, ... , ξ_{1r}, ξ_{21}, ... , $\xi_{2\sigma}$), equate to zero the resulting partial derivatives, and solve this equation system.

1. Maximum Likelihood:
   b. Obtain Maximum likelihood estimate of ρ_{PC}

      for example, the partial derivative of ρ is

$$\frac{\partial Ln(L)}{\partial \rho} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{P_{ij}}\Big[\phi(\xi_{1i},\xi_{2j};\rho) - \phi(\xi_{1i-1},\xi_{2j};\rho)\Big] - \phi(\xi_{1i},\xi_{2j-1};\rho) + \phi(\xi_{1i-1},\xi_{2j};\rho)$$

   c. Solve iteratively using the L information matrix … Computationally demanding!

2. Two-step algorithm
   a. Fit univariate standard normal distributions to the marginal distributions of X_1 and X_2 separately and estimate thresholds
   b. Solve the $Ln(L)$ for ρ only

4.

Where can we calculate polychoric correlations:

The R polycor package

Where can we get Polychoric correlations?

Commercial software:

- PRELIS/ LISREL
- STATA 8 (Polychoric program by Stas Kolenikov)
- SAS (Polychoric Macro from SAS)

Open Source software

- R (Polycor package by John Fox)

The R *polycor* package

'Polychoric and Polyserials correlations package' by John Fox (jfox@mcmaster.ca)

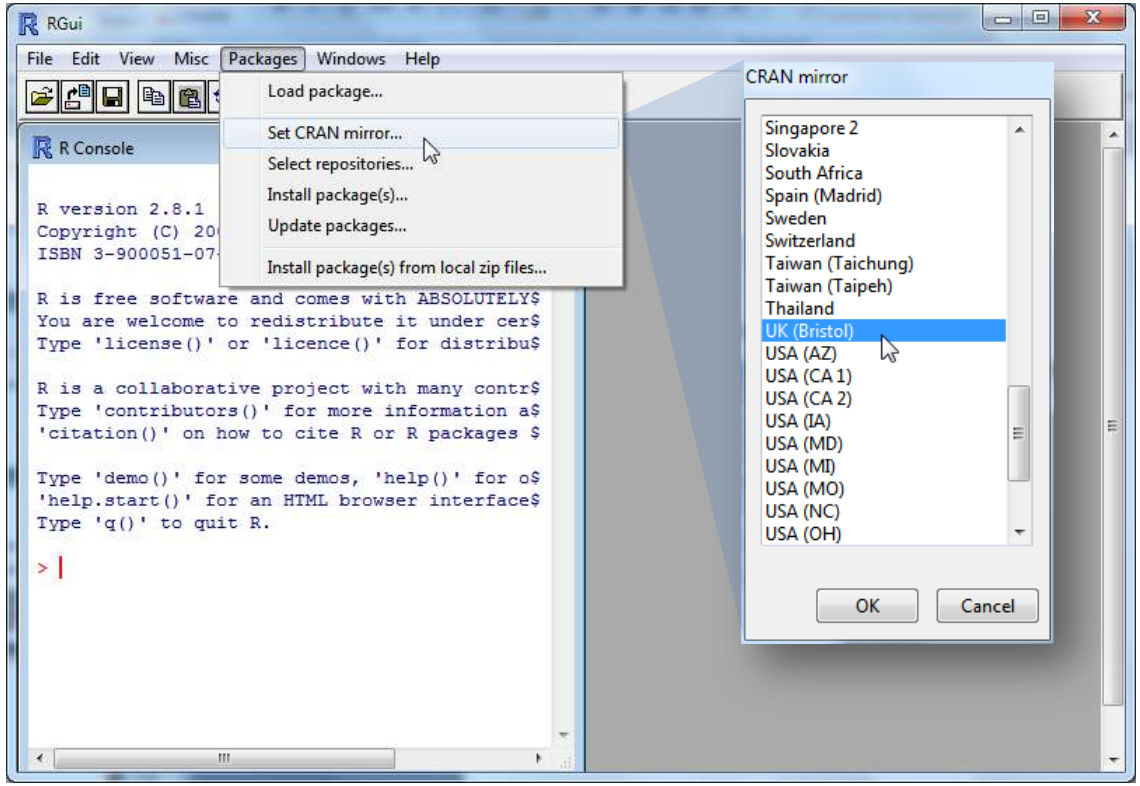
Freely available at CRAN's R repositories

Computes polychoric and polyserial correlations by quick "two-step" methods or ML, optionally with standard errors.

Three programs:

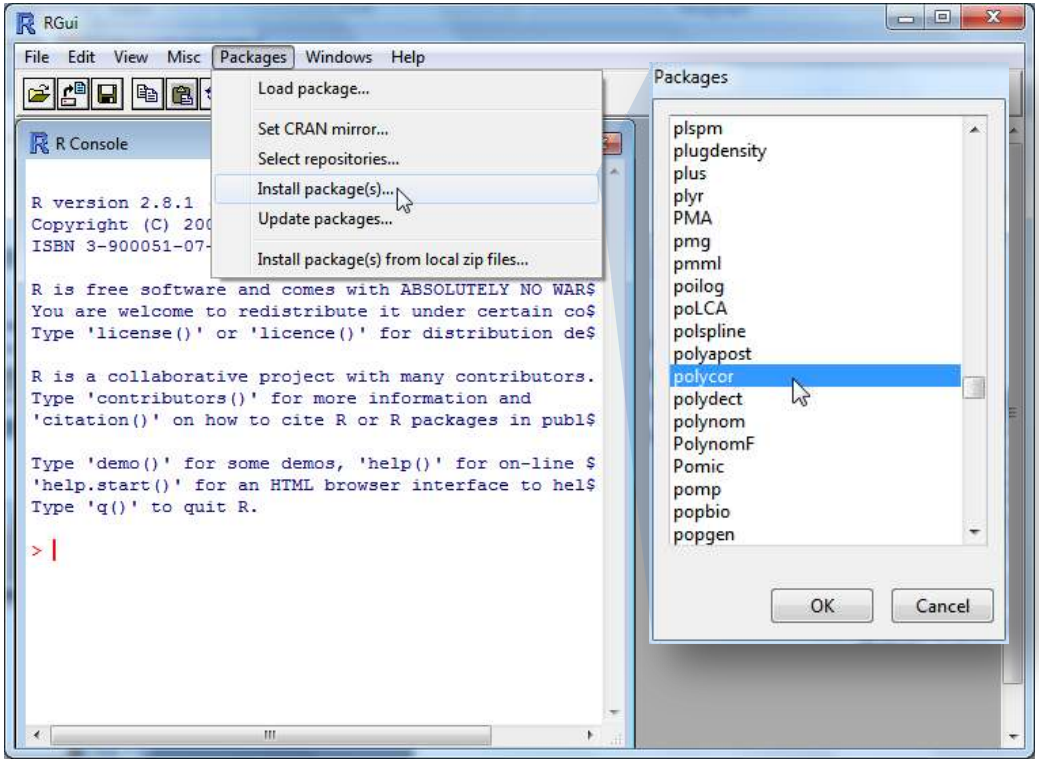
- ***HetCor*** : Computes a heterogeneous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables.
- *Polychor* : Computes the polychoric correlation (and its standard error) between two ordinal variables or from their contingency table.
- *Polyserial*: Computes the polyserial correlation (and its standard error) between a quantitative variable and an ordinal variables.

A. Install Polycor package in R
   - Select CRAN mirror



A. Install Polycor package in R

- Install packages…
- Select ‘polycor’
- OK



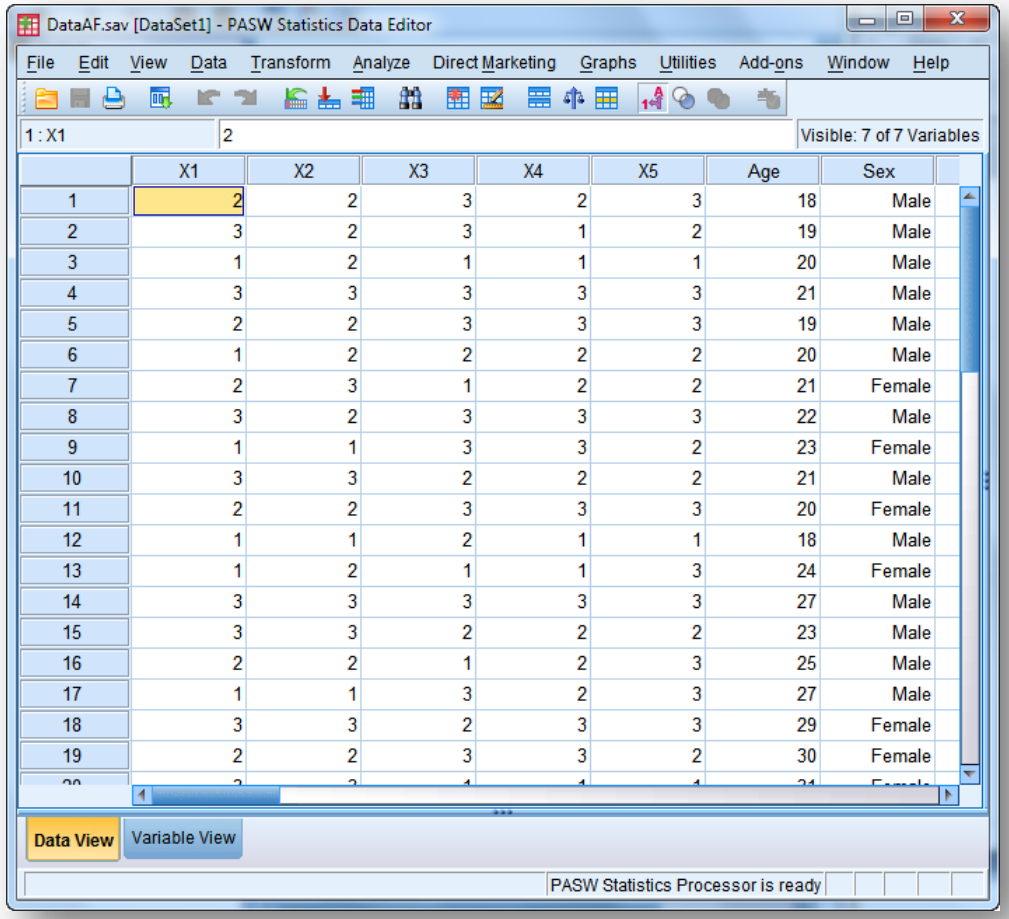
B. Run *hetcor* program

We will use an SPSS data file named 'DataAF.sav':

5 Likert type items (X1,…,X5)

2 Scale variable (Age, SchoolYrs)

1 nominal variable (Sex)



	X1	X2	X3	X4	X5	Age	Sex
1	2	2	3	2	3	18	Male
2	3	2	3	1	2	19	Male
3	1	2	1	1	1	20	Male
4	3	3	3	3	3	21	Male
5	2	2	3	3	3	19	Male
6	1	2	2	2	2	20	Male
7	2	3	1	2	2	21	Female
8	3	2	3	3	3	22	Male
9	1	1	3	3	2	23	Female
10	3	3	2	2	2	21	Male
11	2	2	3	3	3	20	Female
12	1	1	2	1	1	18	Male
13	1	2	1	1	3	24	Female
14	3	3	3	3	3	27	Male
15	3	3	2	2	2	23	Male
16	2	2	1	2	3	25	Male
17	1	1	3	2	3	27	Male
18	3	3	2	3	3	29	Female
19	2	2	3	3	2	30	Female

To import PASW (*.sav) data to R one can either:

1. Within PASW: Export data file to '*.Dat' format and read it from within R

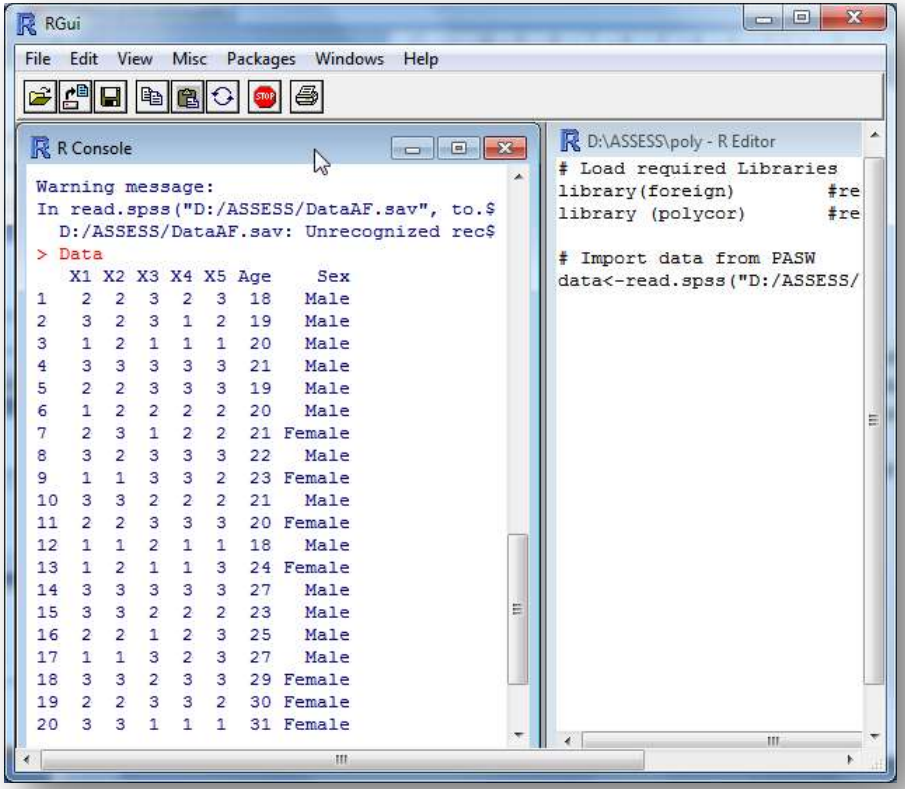
   ```
   > Data<-read.table("D:/ASSESS/DataAF.dat", header=TRUE)
   ```

2. Use `library(foreign)` to import PASW files

   ```
   > library (foreign)
   > data<-read.spss("D:/ASSESS/DataAF.sav", to.data.frame = TRUE)
   ```

   Make sure the data file was imported ok:

   ```
   > Data
   ```



1. Load Library polycor

```
> Library(polycor)
```

2. Assign correct measurement scale to variables (even if you already did it in PASW)

```
> Data$X1<-ordered(Data[,1]) # convert var. to ordered factor
> Data$X2<-ordered(Data[,2])
( )
> Data$X5<-ordered (Data[,5])
> Data$Age<-as.numeric(Data[,6])# convert var. to numeric
> Data$Sex<-factor(Data[,7]) # convert var. to nominal factor
> SchoolYrs<-as.numeric(Data[,8]
```

Notes:

in R "<-" is the same as "="

Data frames are matrix with rows and columns e.g. `Data [2,1]` is the element in the 2nd row and first column; `Data[ ,1]` is the first column which can be evoked as `Data$columnname`

3. Calculate polychoric, polyserial or pearson correlation accordingly to var. measurment:

```
> R<-hetcor(Data, ML=FALSE, std.err=TRUE)
> print (R)
```

Notes:

Options are:

`ML=TRUE` (Default)

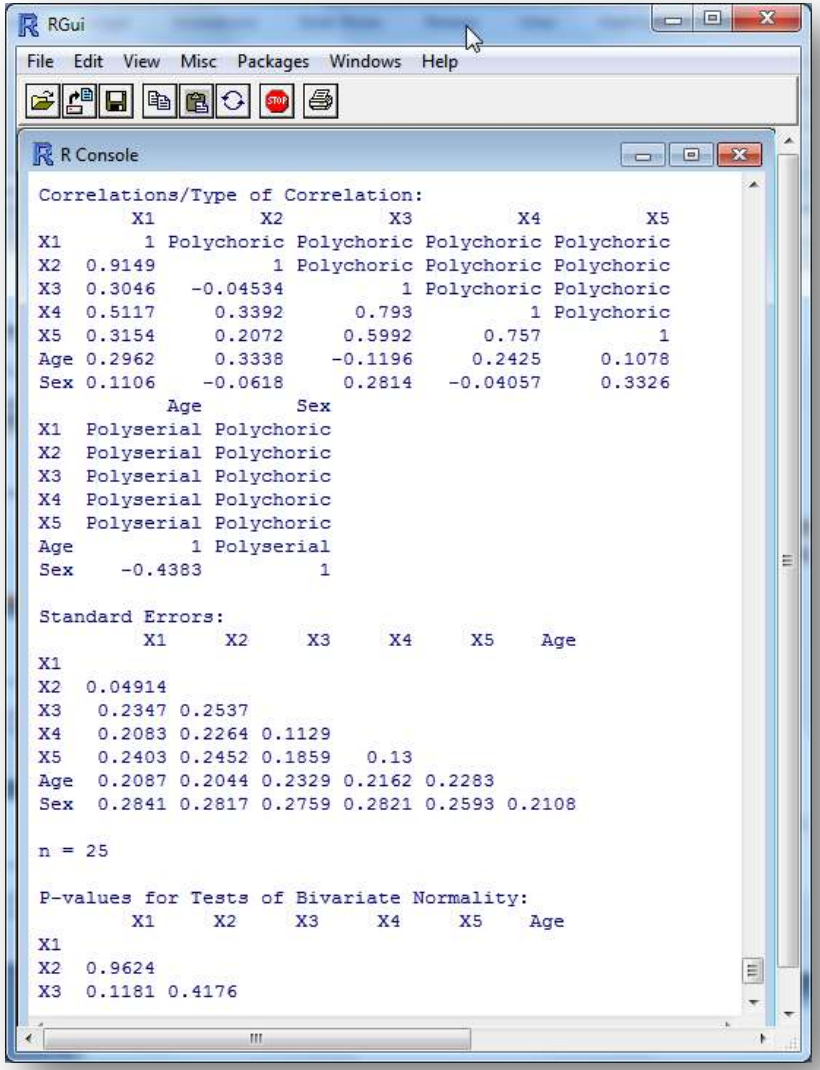
`ML=FALSE` (two-step algorithm)

`std.err=TRUE` (Default)

`std.err=FALSE` (does not show SE)

R is case sensitive:

'`Print`' is not the same as '`print`'



4. Calculate any other correlations

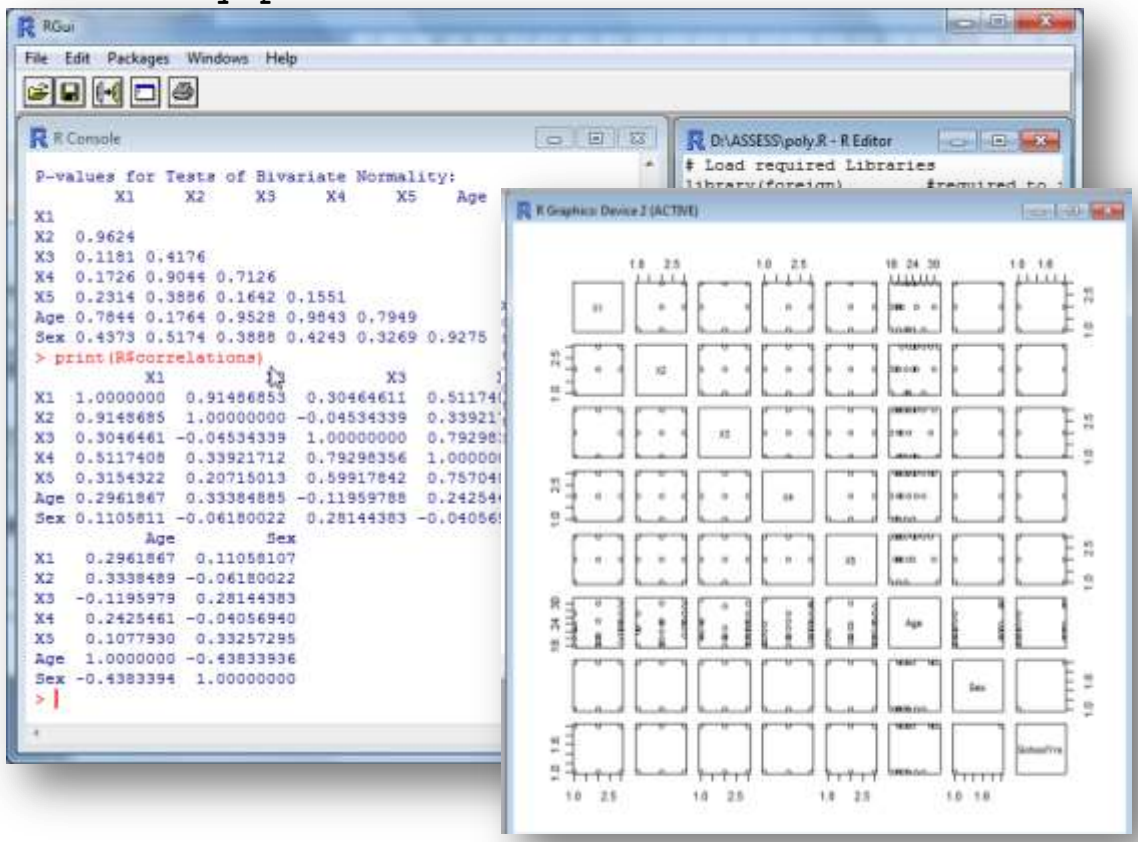
```
> RX1X2<-hetcor(Data$X1,Data$X2, ML=FALSE) #Correlation between
                                           #X1 and X2
> print(RX1X2)
```

5. Extract only a fraction of the R object (**hetcor** produces a List)

```
> str(R)                    #see structure of R object
> print(R$correlations)      #only prints the correlation matrix
```

6. Print scatter matrix of data

```
> pairs(Data)
```



Too much typing?

… only the first time if one saves a script *.R file …

- File
- Save (CTRL+S)

D:\ASSESS\poly.R - R Editor

```
# Load required Libraries
library(foreign)        #required to import PASW .sav files
library (polycor)       #required for the polychoric correlation

# Import data from PASW
Data<-read.spss("D:/ASSESS/DataAF.sav", to.data.frame = TRUE)

# Assign correct measurment levels to variables
Data$X1<-ordered(Data[,1])                # convert variable to ordered factor
Data$X2<-ordered(Data[,2])
Data$X3<-ordered(Data[,3])
Data$X4<-ordered (Data[,4])
Data$X5<-ordered (Data[,5])
Data$Age<-as.numeric(Data[,6])            # convert variable to numeric
Data$SchoolYrs<-as.numeric(Data[,7])
Data$Sex<-factor(Data[,8])                # convert variable to nominal factor

#Calculates Heterelogous correlations
R<-hetcor(Data, ML=FALSE, std.err=TRUE)   #Estimates with ML; if ML=FALSE,estin
print(R)                                  #Note: R is case sensitive: 'Print' i

RX1X2<-hetcor(Data$X1,Data$X2, ML=FALSE)        #Correlation between X1 and X2

str(R)                                    #see structure of R object
print(R$correlations)                     #only prints the correlation matrix
```

5.

How about PASW Statistics?

The PASW Statistics - R Integration package

How about PASW Statistics?

There is no poychoric or polyserial correlations in PASW Statistics (up to v18).

However, SPSS v16 or higher allows one to include R code into SPSS Syntax.

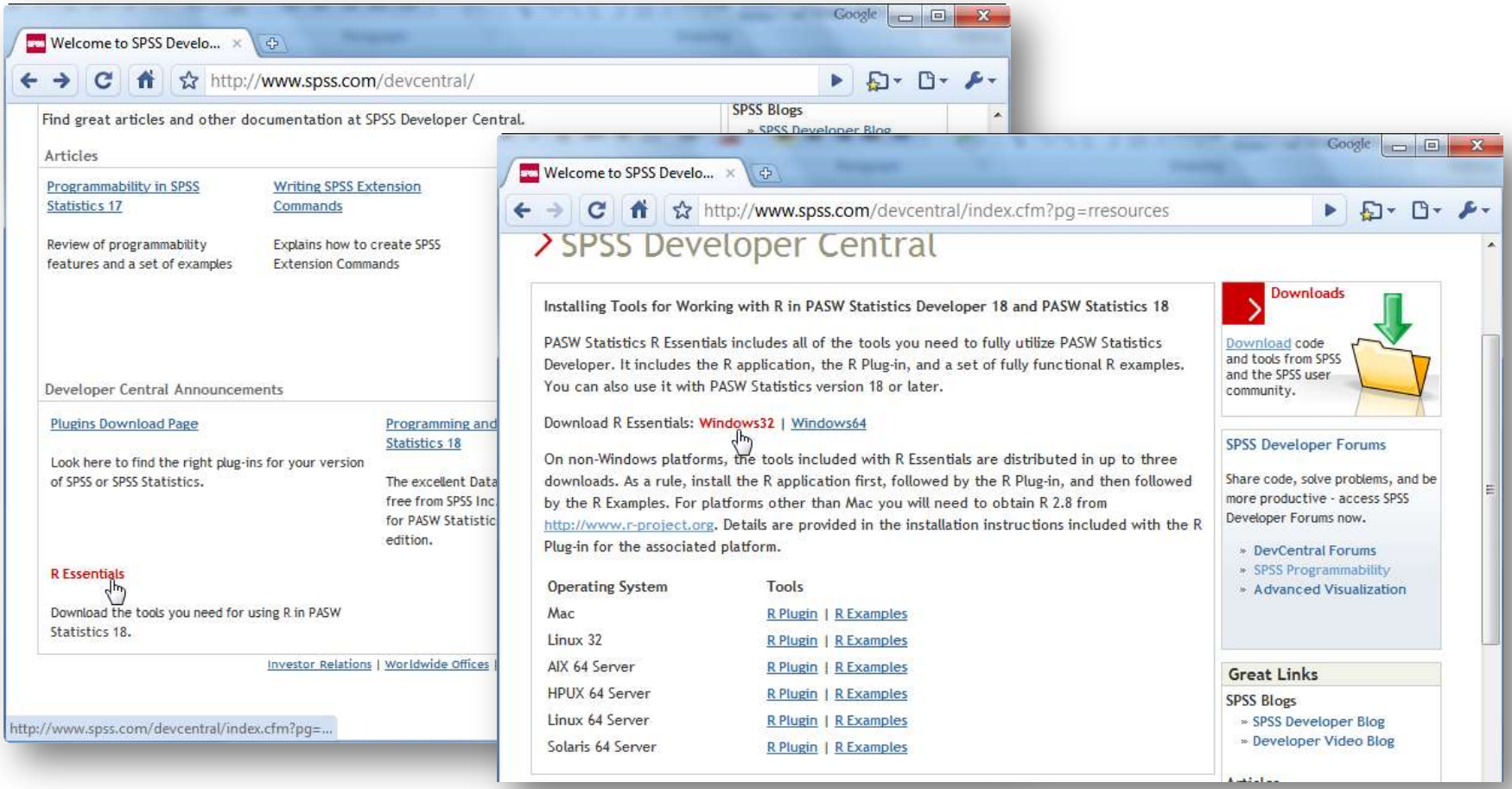
PASW18 (just released) has an 'R Essentials' package:

- PASW :R integration plug-in
- R 2.8.1
- Several examples of R libraries portrayed to PASW Statistics 18

The PASW-R integration plug-in and its R package provides functions for:

1. Read case data from the active dataset into R
2. Get variable information in the active dataset
3. Get output results (via OMS) from PASW into R
4. Write R results back to PASW output viewer
5. Display R graphics in the PASW output viewer

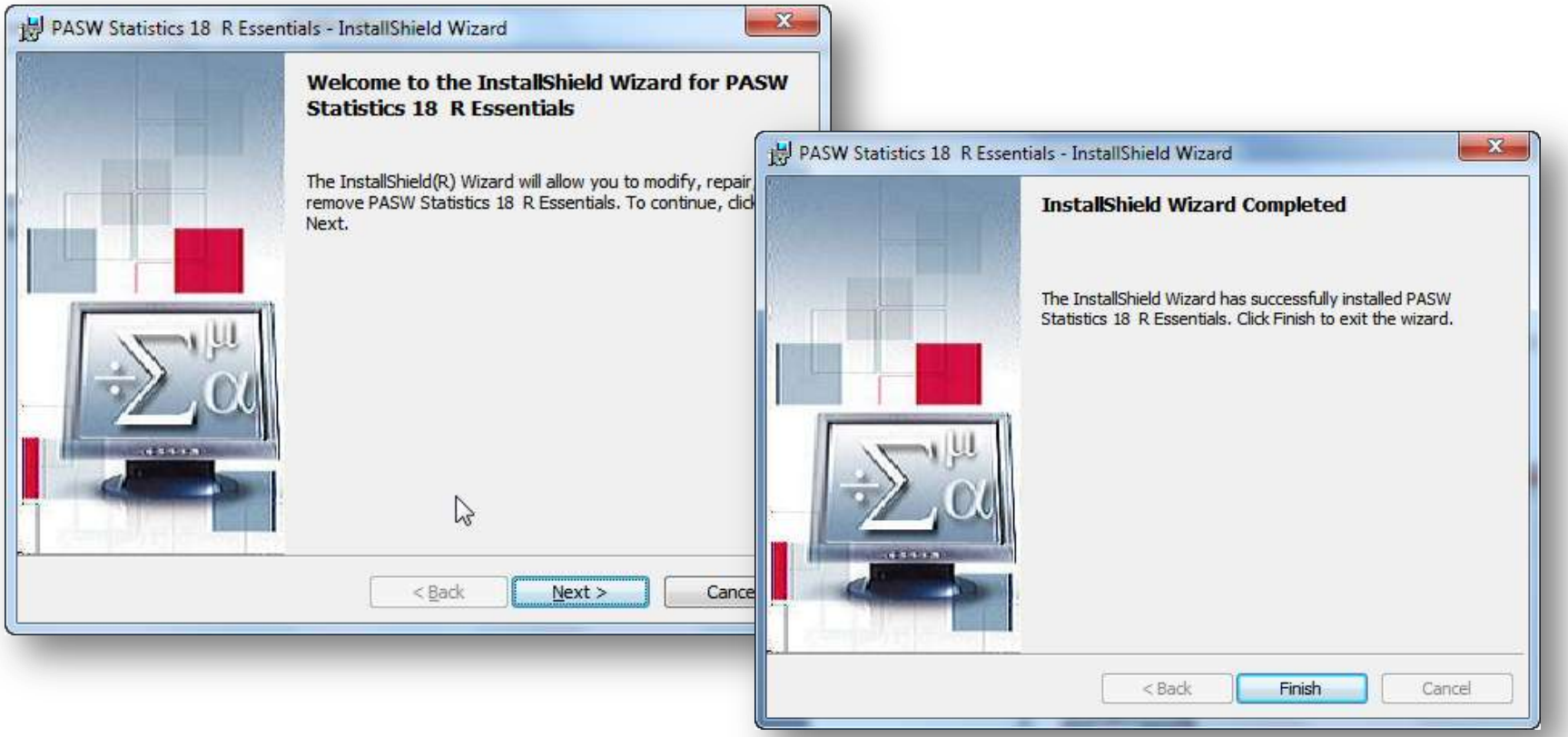
- Browse to http://www.spss.com/devcentral/
- Register a free account
- Navigate to '**`R Essentials`**' (bottom of the page)
- Download R essentials for Windows32 or Windows64 (if other OS, see below)
- (accept the legal disclaimers) Install **`R essentials`**



R essentials

`R essentials` will install:

- R 2.8.1 application (R is actually in version 2.9.2, but for compatibility proposes its better to keep 2.8.1)
- PASW Statistics R Plug-in
- PASW Statistics R examples with CDB and syntax extensions

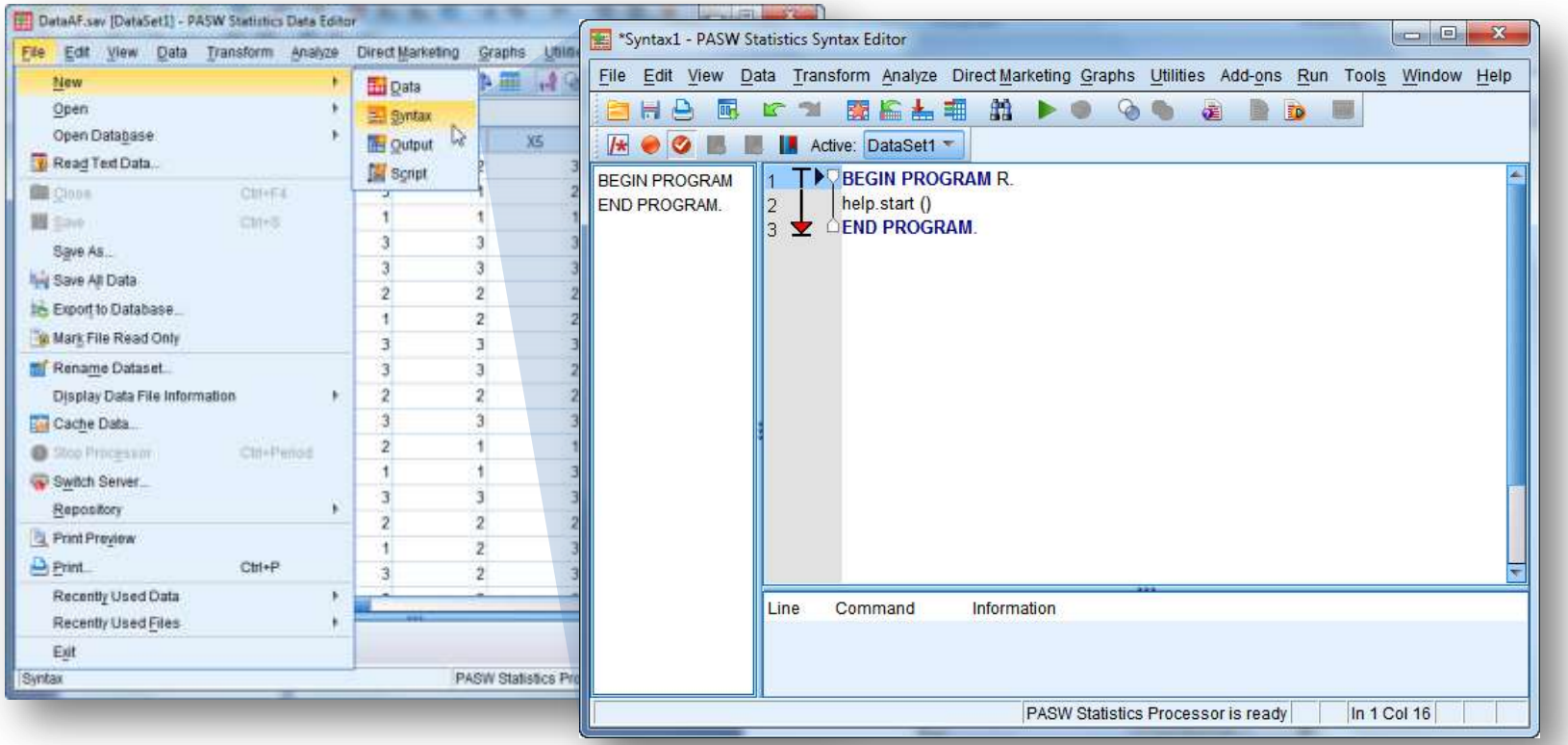


R code into PASW Statistics

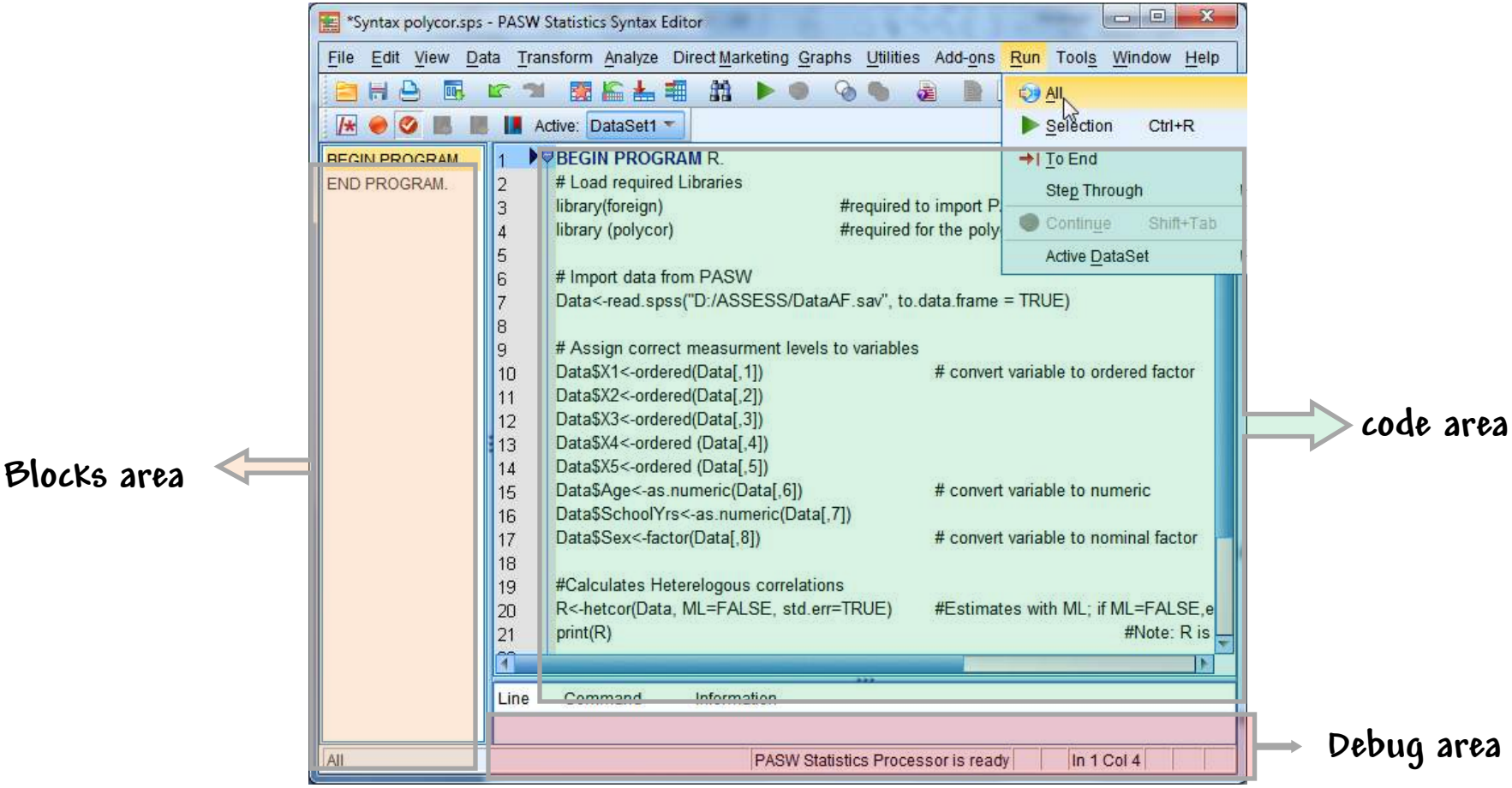
In PASW Statistics:

- File ▸ New ▸ Syntax
- R code must be enclosed in between (syntax block)

```
BEGIN PROGRAM R.
( )
END PROGRAM. */ don't forget these "."
```

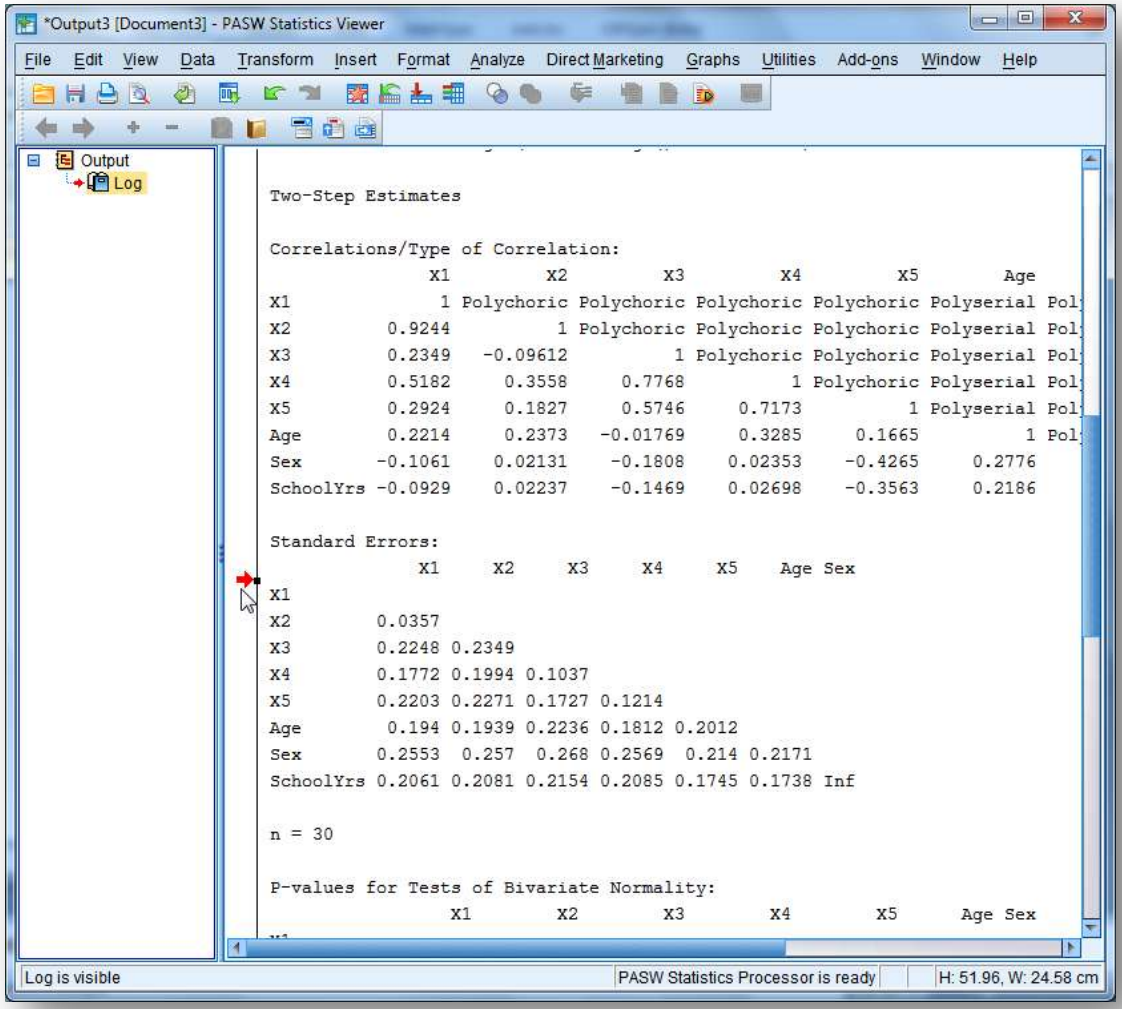


Can just paste the code we did before:



- Run ▸ All

'R output' will be parsed as a text block into 'PASW output viewer':



But, we should:

1. Get variables directly from PASW Statistics active data set
2. Route R output to PASW Statistics output tables

1. To get data from active PASW Statistics:

```
>Data<-spssdata.GetDataFromSPSS(variables="X1 X2 X3 X4 X5 Age Sex SchoolYrs")
                                # Reads the variables from SPSS active datasheet
```

To get variable information from active PASW Statistics:

```
>vardict<-spssdictionary.GetDictionaryFromSPSS(variables="X1 X2 X3 X4 X5 Age Sex
   SchoolYrs")                          # Gets variables definition from
                                        # PASW: It is fundamental that the
                                        # right measurement scale is used
```

Missings in the Data? Handle them in R:

```
>is.na(Data)<-is.na(Data)
>Data<-na.omit(Data)
```

Or, as per v.18:

```
>Data<-spssdata.GetDataFromSPSS(variables="X1 X2 X3 X4 X5 Age Sex SchoolYrs“,
   missingValueToNA=TRUE)
```

Pass the variable's information (measurement level) to R:

Loop for i=1 to length of Data frame

Check PASW measurement level

Convert vectors into numeric, ordered or factor R vectors accordingly to measurement level

```
for (i in 1:length(Data)) {
 if (vardict["varMeasurementLevel",i]=="scale") Data[,i]<-Data[,i]
  else
 if (vardict["varMeasurementLevel",i]=="nominal") Data[,i]<-factor(Data[,i])
  else
 if (vardict["varMeasurementLevel",i]=="ordinal") Data[,i]<-ordered(Data[,i])
}
```

OR, for PASW18 or higher:

`spssdata.GetDataFromSPSS("x1 x2 x3", factorMode="labels")` Will get the correct factors (nominal or ordinal into R with value labels)

`spssdata.GetDataFromSPSS("x1 x2 x3", factorMode="levels")` Will get the correct factors (nominal or ordinal into R with values)

Do the calculations:

```
#Calculates Heterogeneous correlations
R<-hetcor(Data, ML=FALSE, std.err=TRUE)
```

2. Route R output to PASW Statistics output tables

Note: Complex R list objects (like the one created by hetcor) cannot be passed directly to PASW Statistics output tables

One has to pass list elements one by one.

To see the structure of the object created in R and the names of its elements do:

```
>str(R)
```

For example, hetcor, produces a list with elements:

`$correlations`	(correlation coeficients)
`$type`	(type of correlations)
`$std.errors`	(std.errors)
`$n`	(sample size)
`$tests`	(p-values for bivariate normality tests)

Write the correlation type table to PASW Statistics output table:

```
spsspivottable.Display(R$types, title="Correlation types")
```

Write the correlations table to PASW Statistics output table:

```
spsspivottable.Display(R$correlations, title="Correlation matrix")
```

Write the Std.Errors table to PASW Statistics output table:

```
spsspivottable.Display(R$std.errors, title="Std. errors")
```

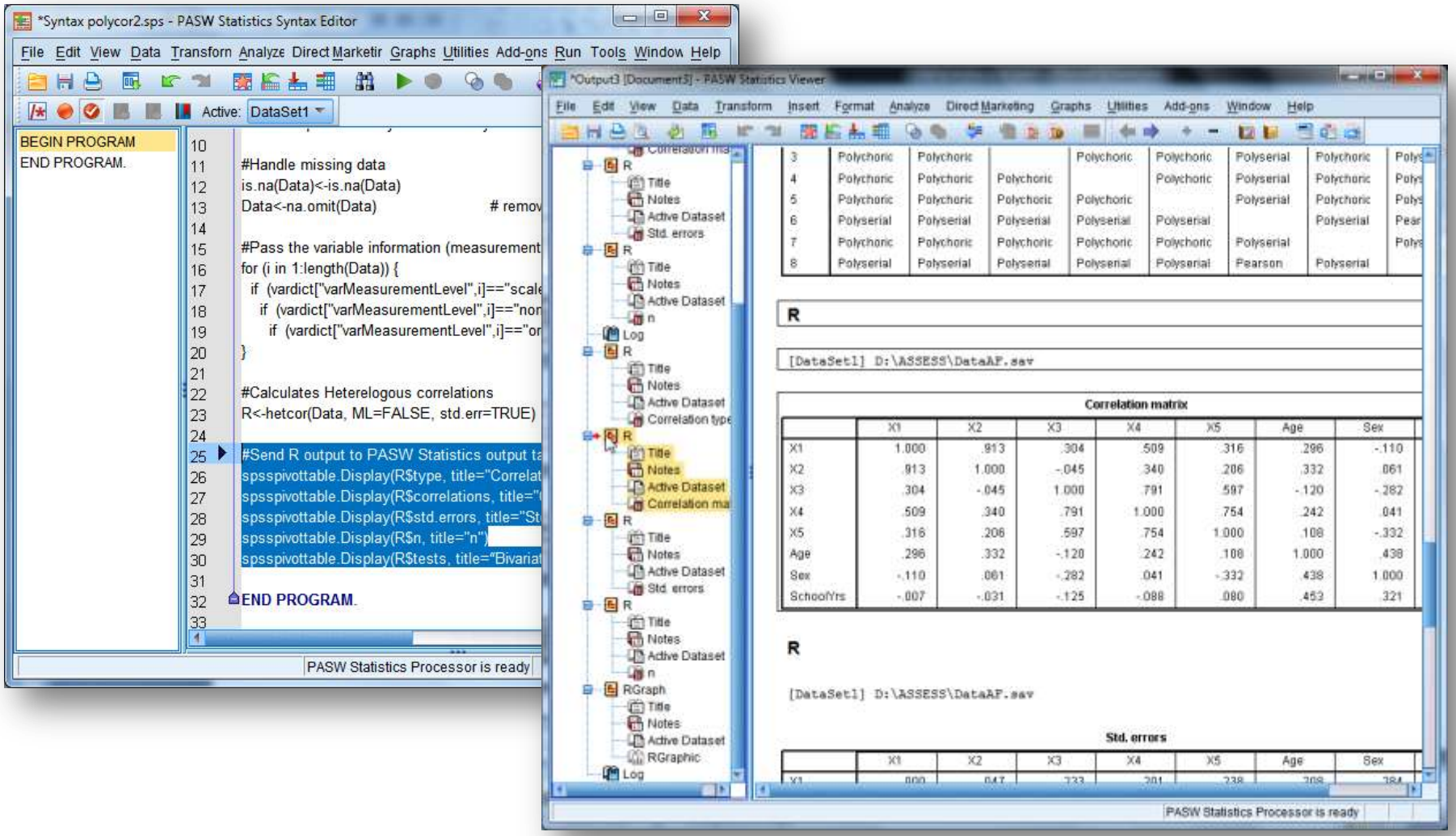
Write the ‘n’ table to PASW Statistics output table:

```
spsspivottable.Display(R$n, title="n")
```

Write the ‘bivariate tests’ table to PASW Statistics output table:

```
spsspivottable.Display(R$tests, title="Bivariate Normality tests")
```

More complex hierarchical tables can be produced with `BasePivotTable()`


*Syntax polycor2.sps - PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet1
BEGIN PROGRAM
END PROGRAM.
#Handle missing data
is.na(Data)<-is.na(Data)
Data<-na.omit(Data)
#Pass the variable information (measurement
for (i in 1:length(Data)) {
#Calculates Heterelogous correlations
R<-hetcor(Data, ML=FALSE, std.err=TRUE)
#Send R output to PASW Statistics output ta
spsspivottable.Display(R$type, title="Correlat
spsspivottable.Display(R$correlations, title="
spsspivottable.Display(R$std.errors, title="St
spsspivottable.Display(R$n, title="n")
spsspivottable.Display(R$tests, title="Bivariat
END PROGRAM.
PASW Statistics Processor is ready
*Output3 [Document3] - PASW Statistics Viewer
File Edit View Data Transform Insert Format Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Title
Notes
Active Dataset
Std. errors
n
Log
Correlation type
Correlation ma
RGraph
RGraphic
R
[DataSet1] D:\ASSESS\DataAF.sav
Correlation matrix
X1 1.000 .913 .304 .509 .316 .296 -.110
X2 .913 1.000 -.045 .340 .206 .332 .061
X3 .304 -.045 1.000 .791 .597 -.120 -.282
X4 .509 .340 .791 1.000 .754 .242 .041
X5 .316 .206 .597 .754 1.000 .108 -.332
Age .296 .332 -.120 .242 .108 1.000 .438
Sex -.110 .061 -.282 .041 -.332 .438 1.000
SchoolYrs -.007 -.031 -.125 -.088 .080 .453 .321
Std. errors
PASW Statistics Processor is ready

Is it necessary to write a new program every time we want to use polycor (or any other R library)?

No!

Implement a Custom Dialog into PASW to:

1. Select variables through familiar PASW statistics menus
2. Produce PASW Statistics regular output

6.

Getting the polycor into PASW menus:

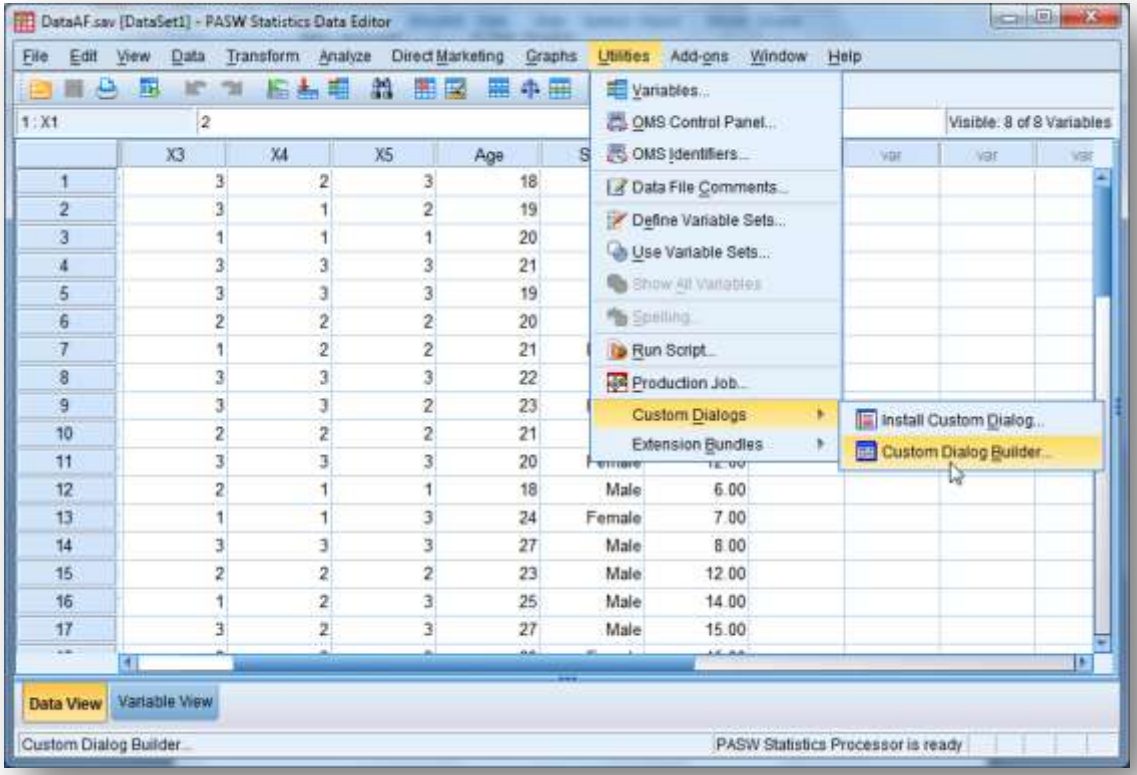
The PASW Custom Dialog Builder

The Custom Dialog Builder

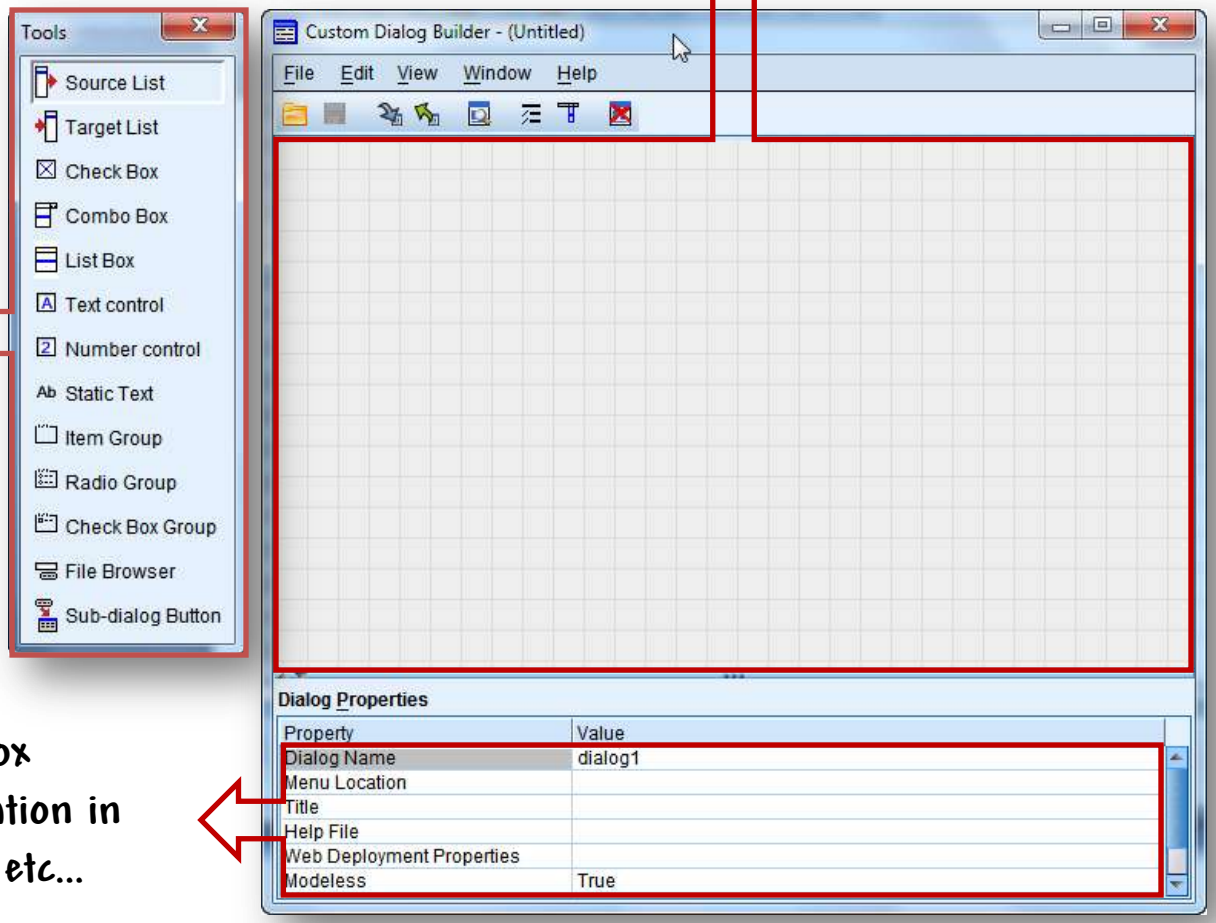
(PASW v17 or higher)

In PASW Statistics:

- Utilities
  - Custom Dialogs
    - Custom Dialog Builder



Menu Layout canvas



Tools box

Properties box
(R code, location in
PASW, Title, etc...

Fill in Dialog properties:

Dialog Name

Title

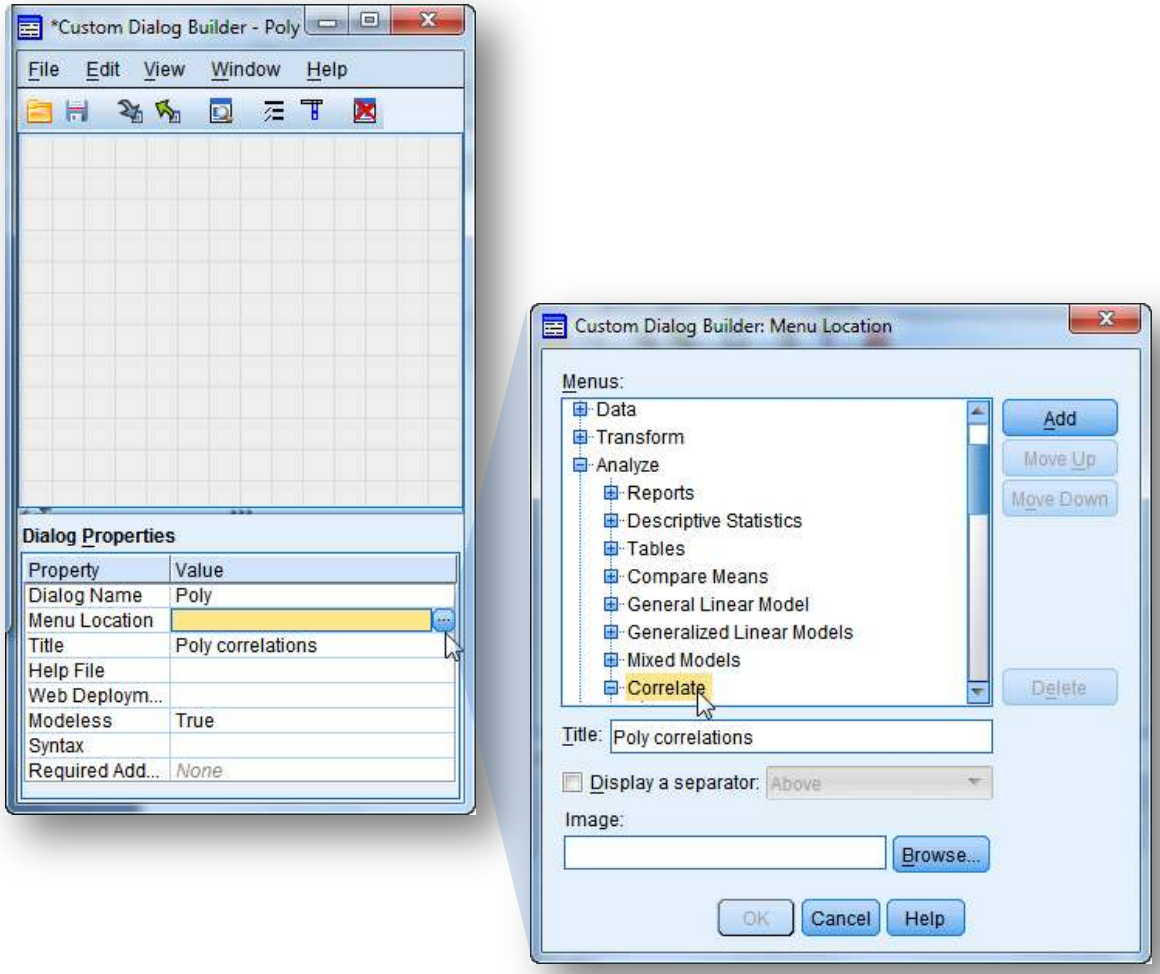
Menu Location: click

Select: Analyze ▸ Correlate

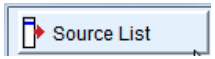
Click Add

Syntax

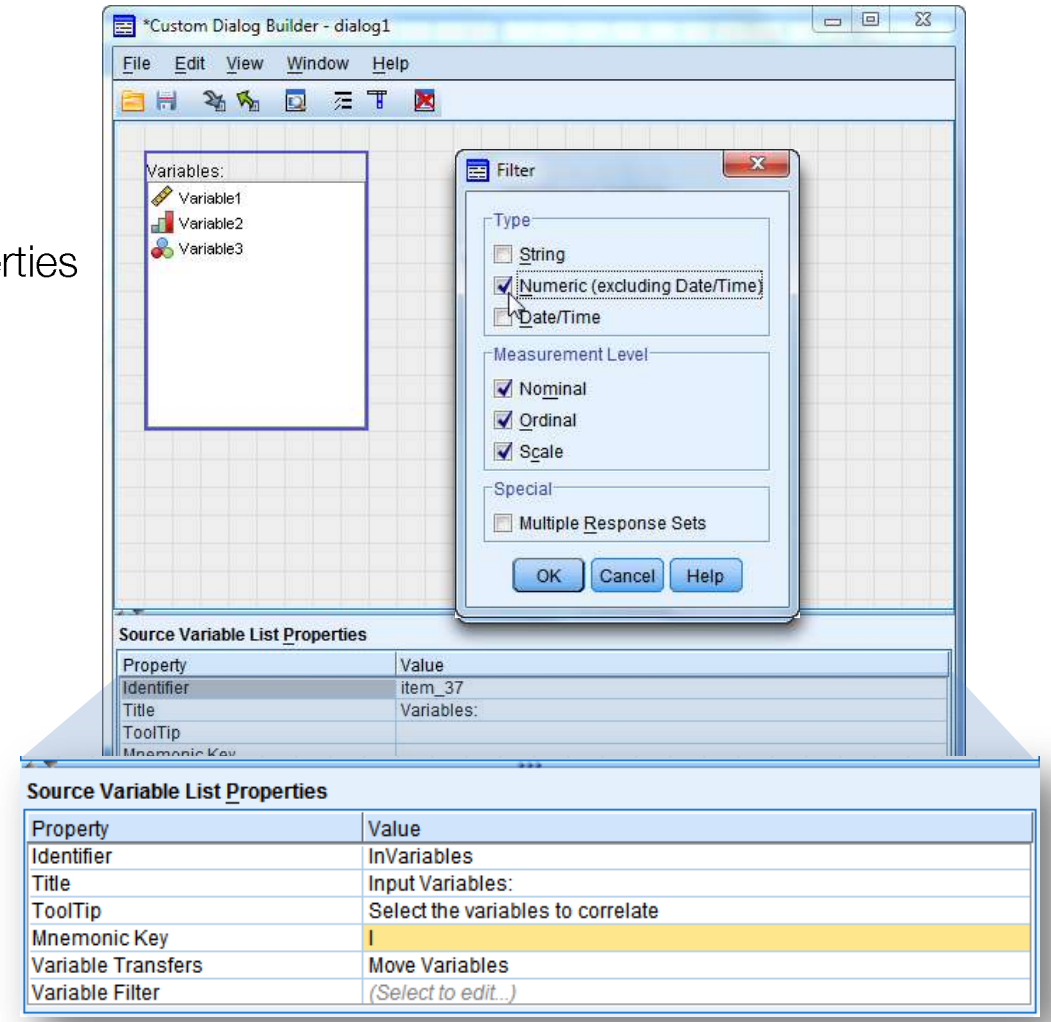
(Leave it alone for now)



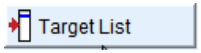
Add Input Variables:

- Select 

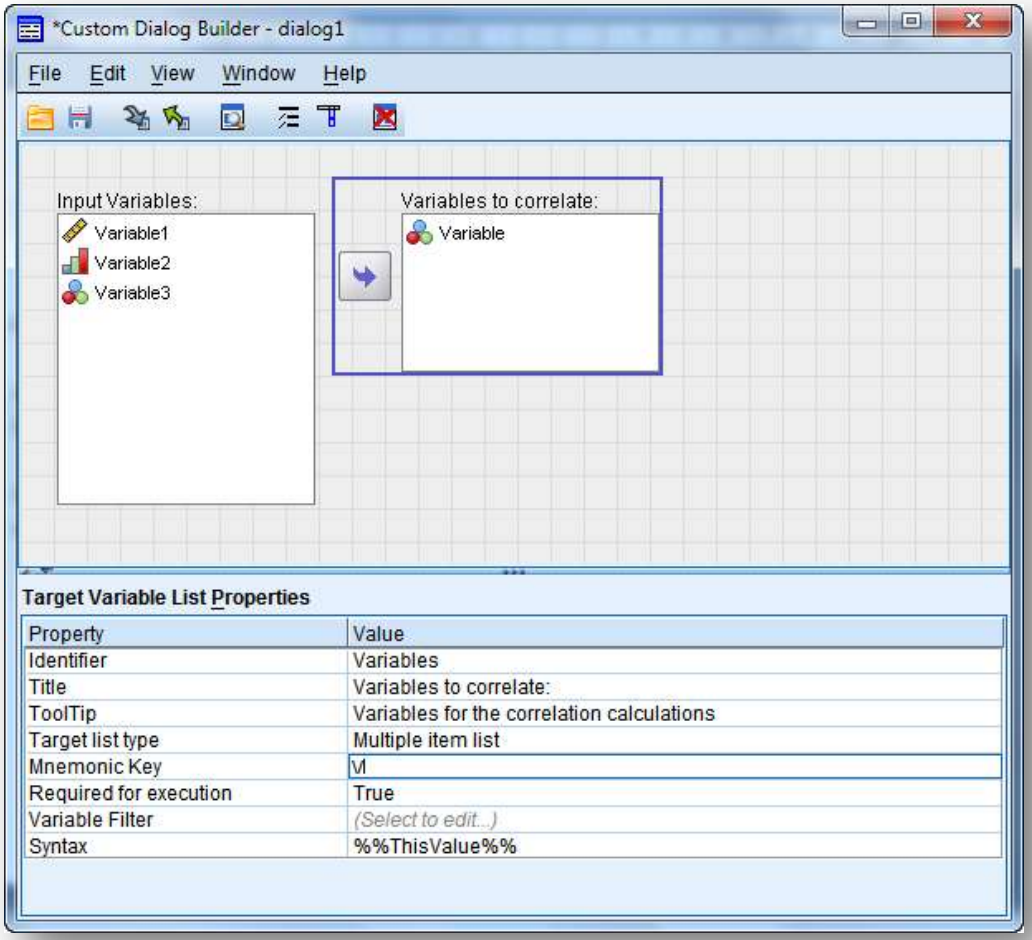
tool
- Drag the Source variables box into drawing canvas
- Double-click to edit properties
- Fill in the 'Source variables List Properties



Add Variables to correlate:

- Select 

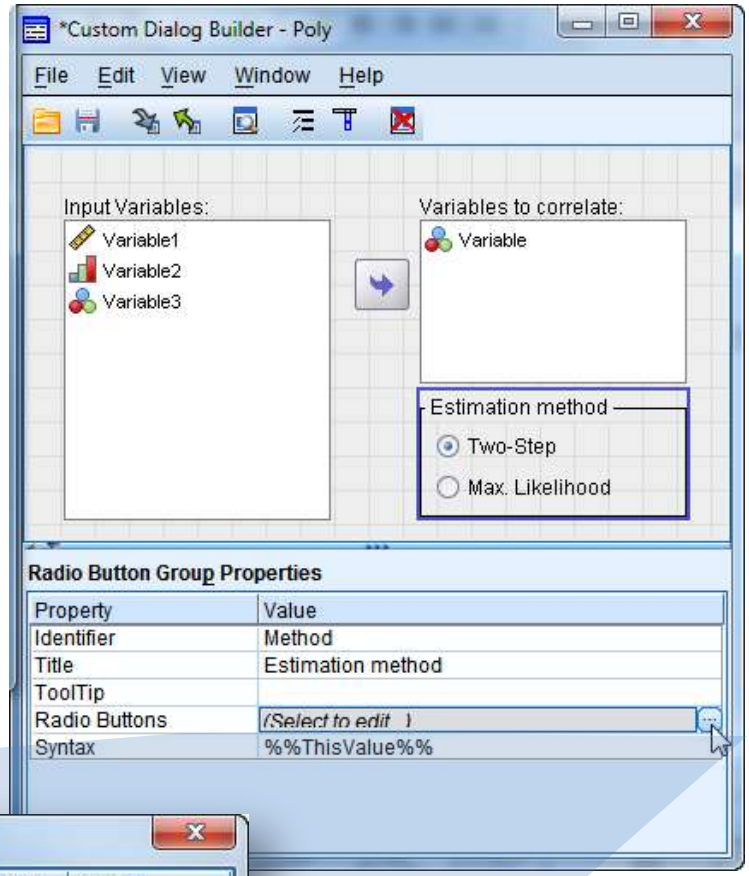
tool
- Drag the Target variables box into drawing canvas
- Double-click to edit properties
- Fill in the 'Source variables List Properties

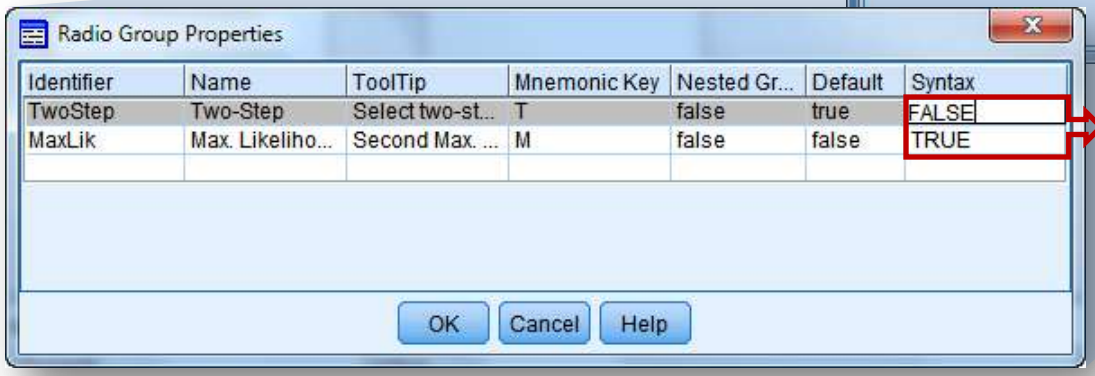


Add Analysis options:

•Method of estimation: ML or Two-step

- Select Check Box Group tool
- Drag into drawing canvas
- Fill in the 'Source variables List Properties
- Click 'Radio Buttons' and change its properties by clicking on

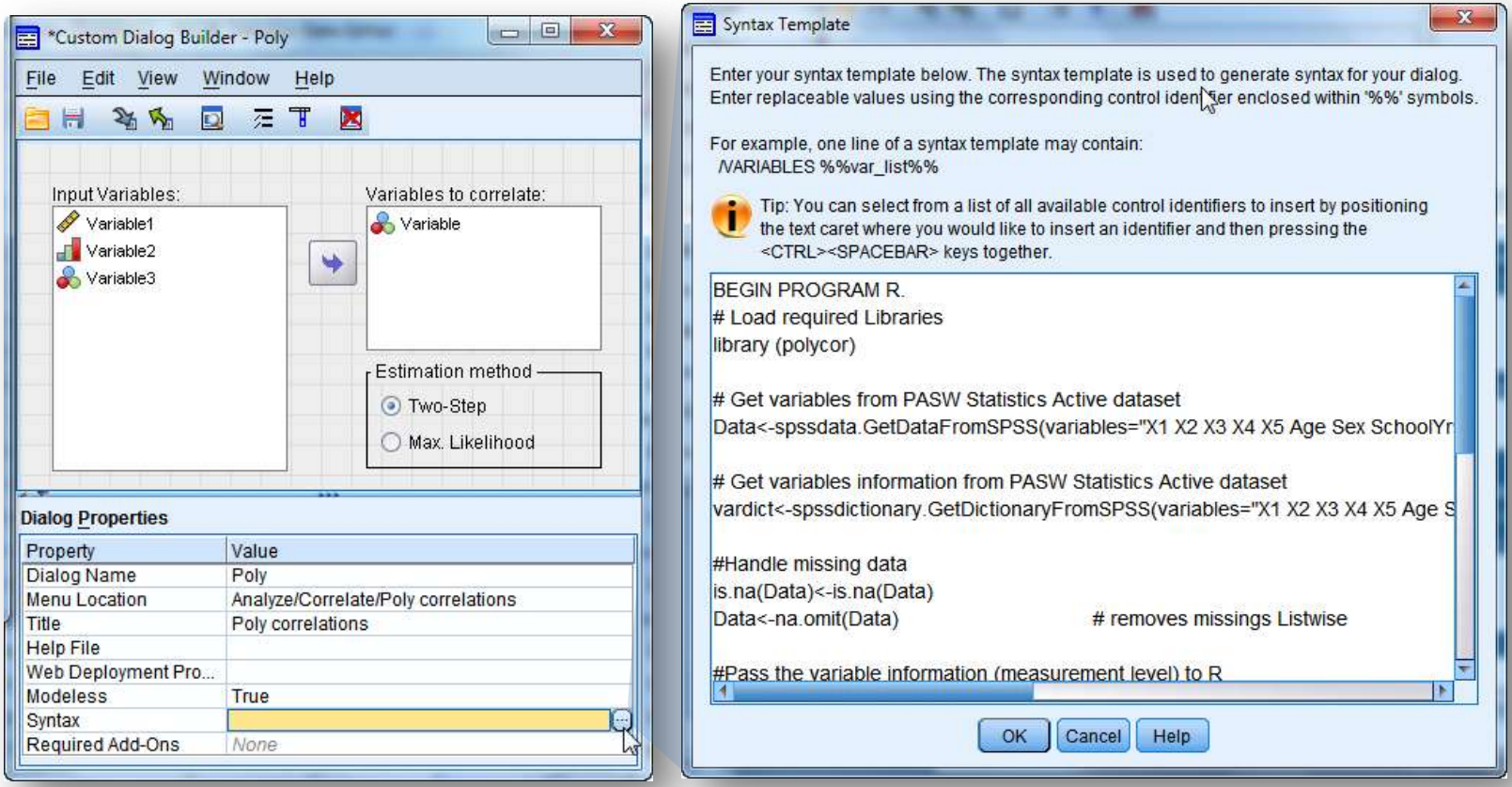




Keywords for R syntax

Get the previously done syntax into the CDB syntax window:

- Syntax ▸
- Copy &Paste syntax into CDB syntax window



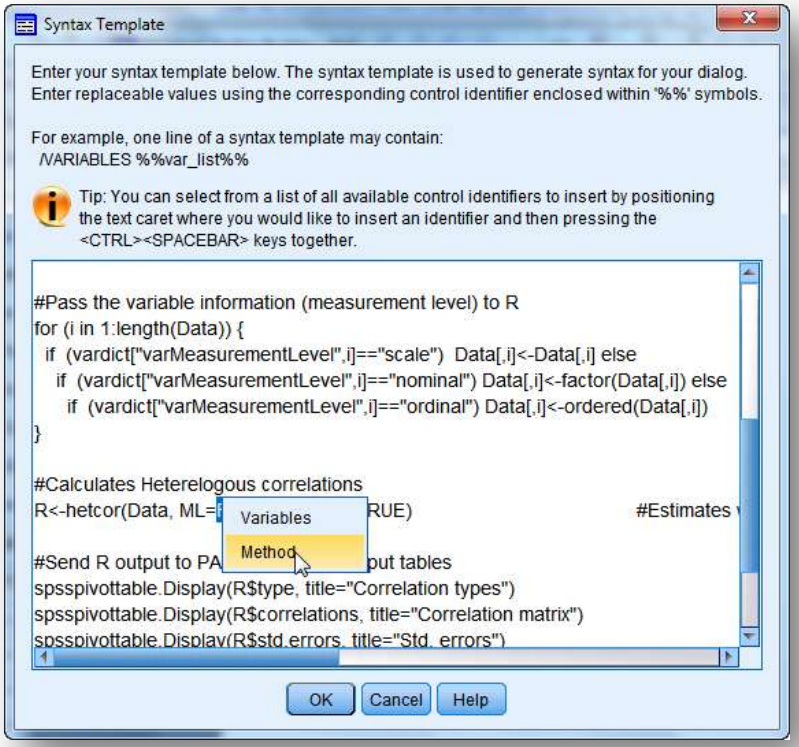
Get the previously done syntax into the CDB syntax window:

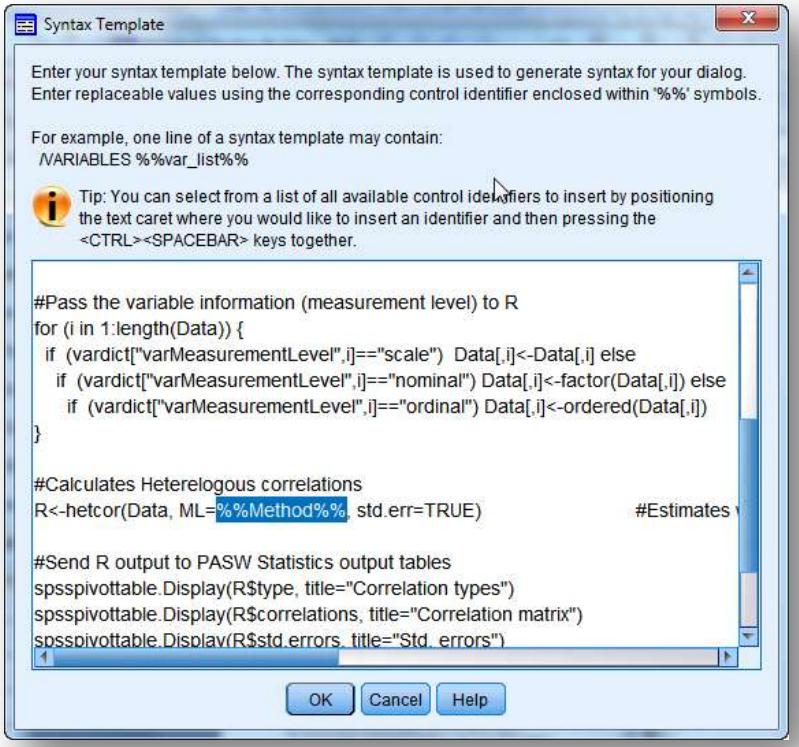
- Correct syntax for CDB specifications:
- CTRL+SPACEBAR to see a list of available CDB created keywords
- Change (both in spps.GetData and spss.GetVariables)

  `"X1 X2 X3 X4 X5 Sex Age SchoolYrs"` to `"%%Variables%%"`

Get the previously done syntax into the CDB syntax window:

- In `hetcor(Data,ML=FALSE, )` Change `FALSE` to `%%Method%%`





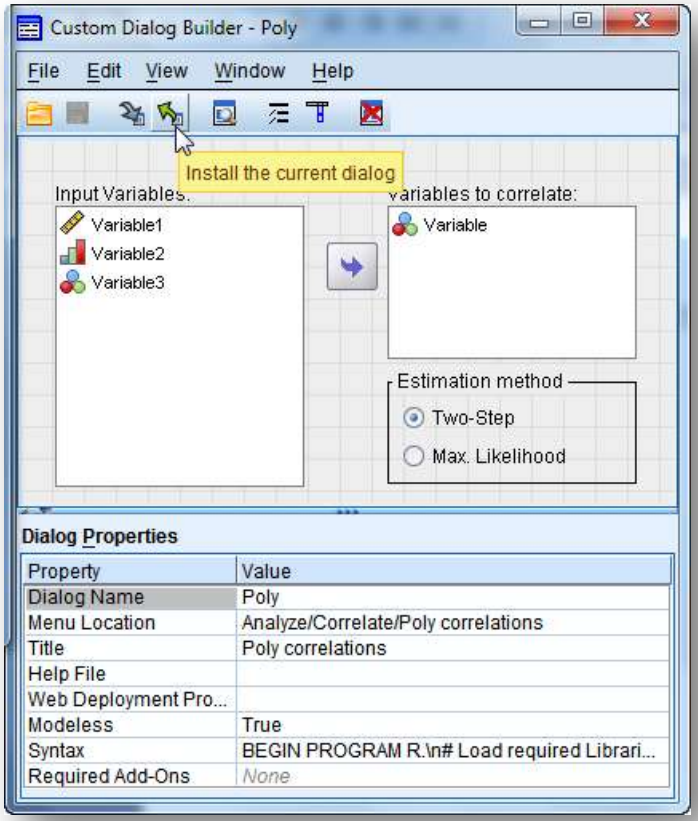
- OK
- File ▸ Save (if you haven't done so…)

Polycor menu implementation with CDB

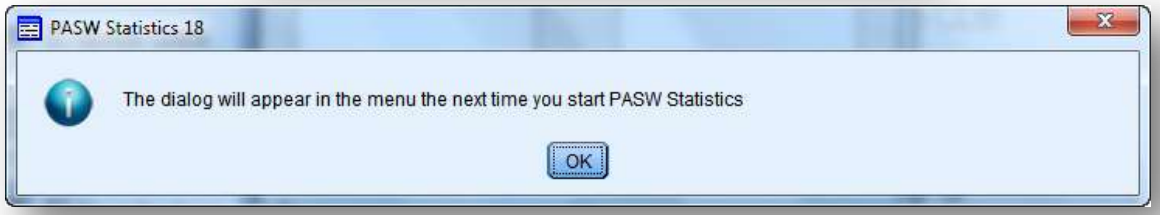
(PASW v17 or higher)

Install the Dialog in PASW Statistics:

▸ File ▸ Install 

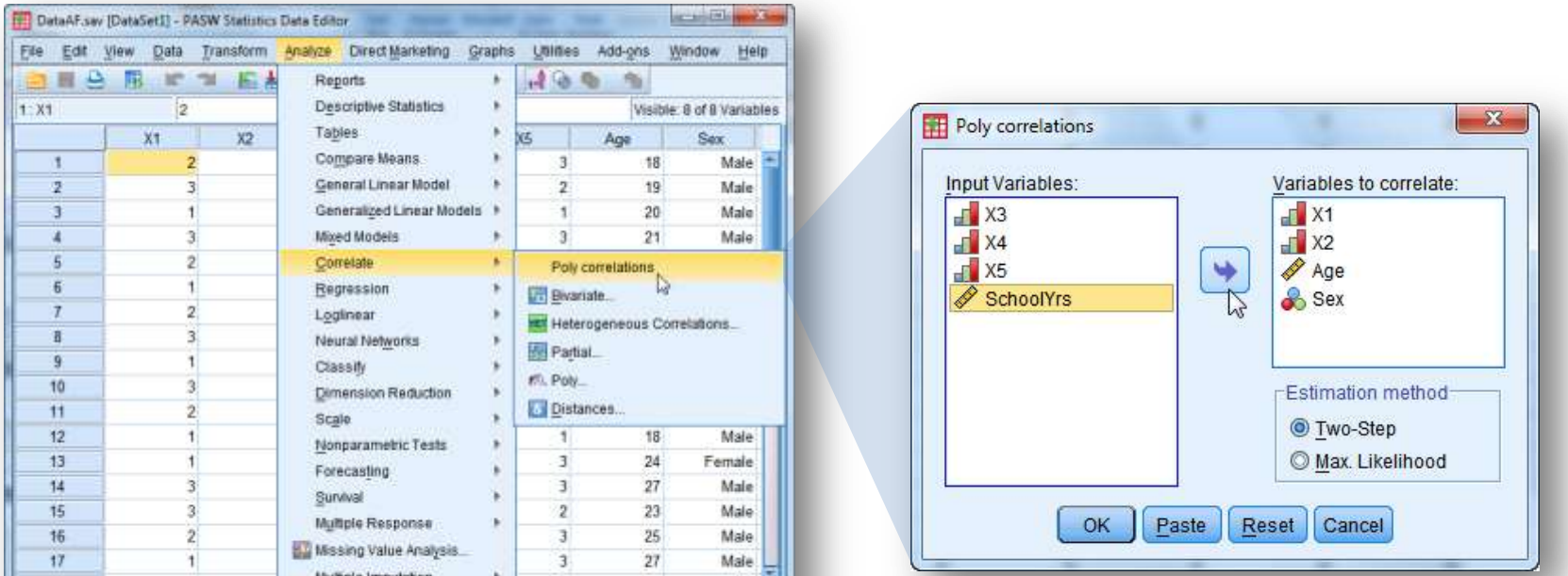


If everything is ok, you should see this:

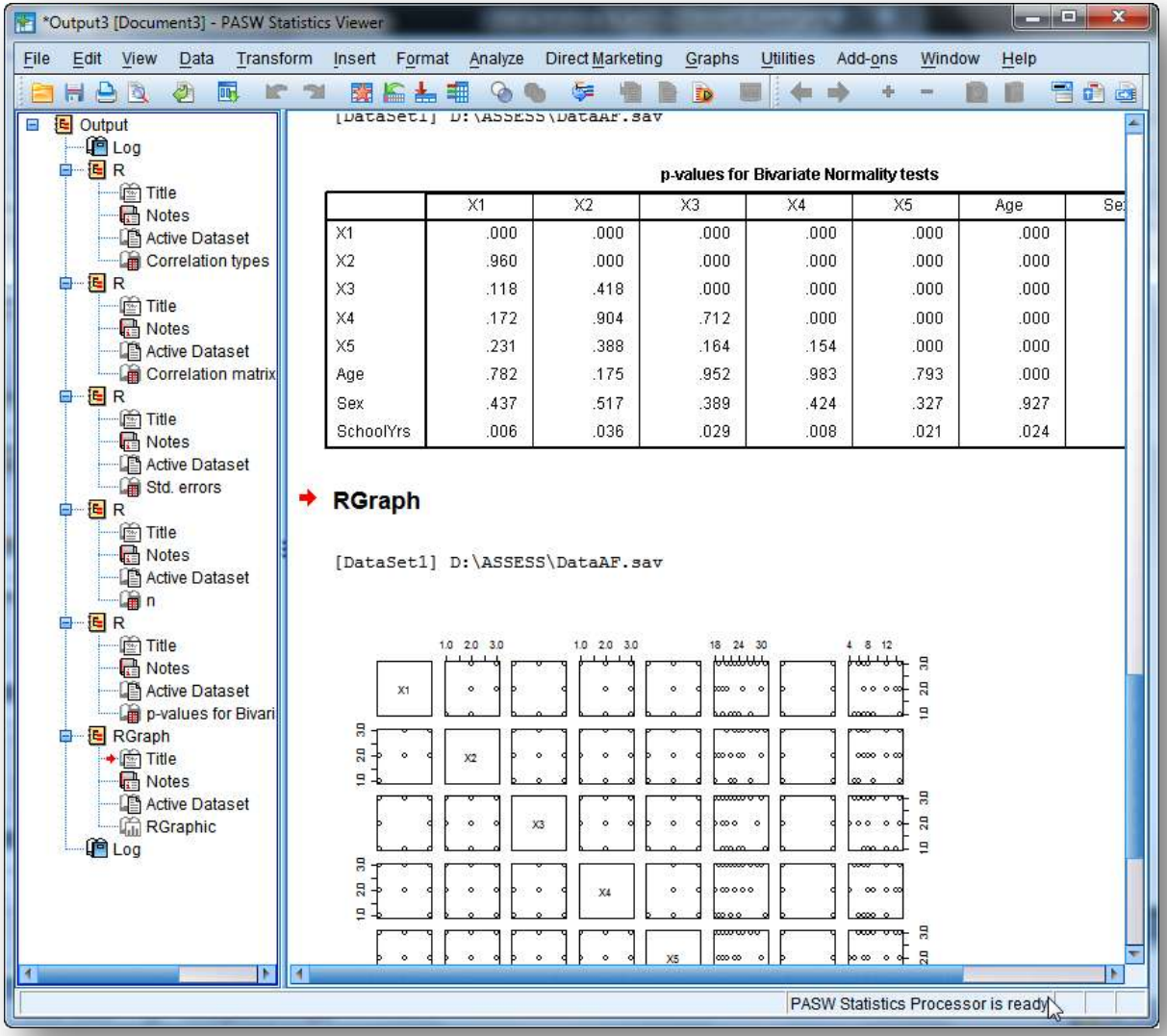


Restart PASW statistics

- Analyze
- Correlate
- Poly correlations:

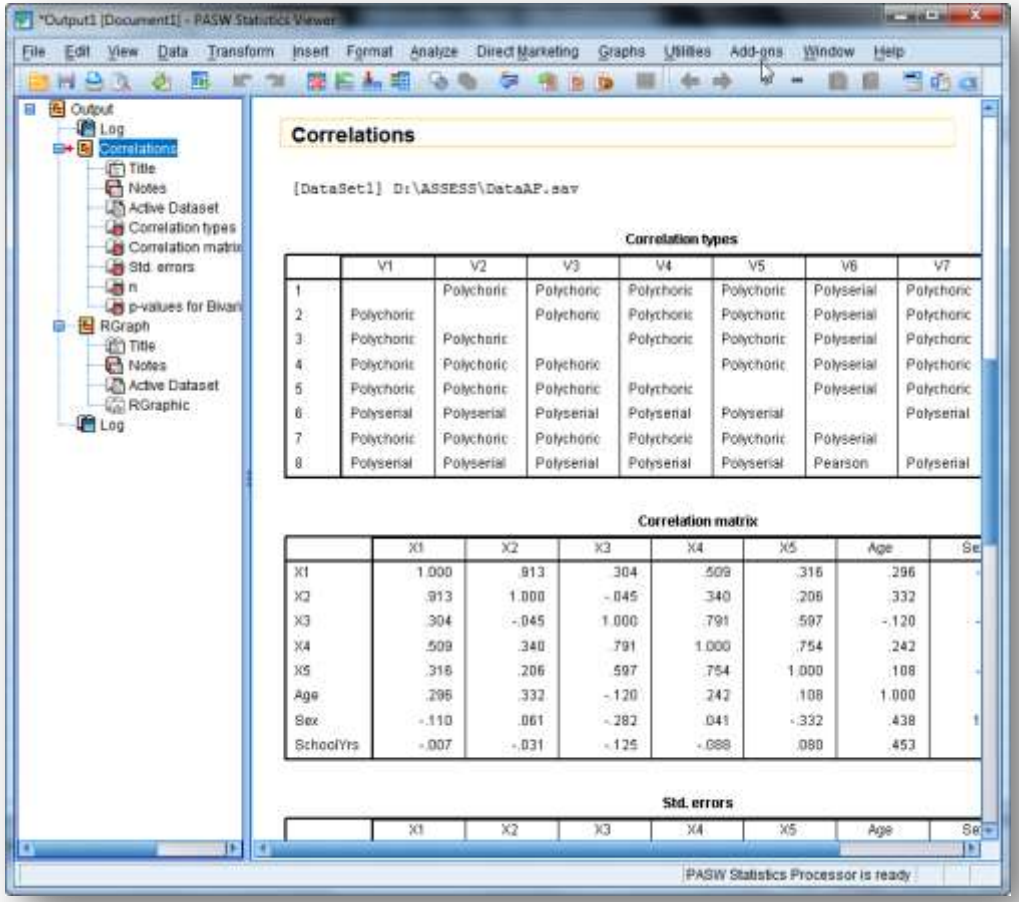


and that is it:



If one wants to create custom output names enclose **spsspivotable** commands within **spsspkg.StartProcedure (“custom name”) … spsspkg.EndProcedure()**

```
spsspkg.StartProcedure ("Correlations")
spsspivottable.Display(R$type, title="Correlation types")
(…)
spsspkg.EndProcedure()
```



With Custom Dialogs one can:

- Create native looking Menus…
- … both for PASW syntax, Python and R (with integration plug-ins)
- Create custom PASW Statistics output tables from the R (or Python) analysis
- Write back analysis results to PASW Statistics active Data sets

```
spssdictionary.SetDictionaryToSPSS ("Results", vardict)
spssdata.SetDataToSPSS ("Results", Data)
```

What's next?

- Save and Distribute easy-to-install Custom Dialogs between PASW end users (see spss.com/devcentral for User and SPSS contributed Custom dialogs)

  (in windows 7, a double click in the *.spd file will install it, automatically, in PASW statistics)
- Convert R code into 'Extension commands' to use R with PASW-like syntax

Need more information?

See SPSS publications (free download at spss.com/devcentral)

- PASW Statistics-R integration package (in R essentials)
- Programming and Data Management for PASW Statistivs18

Questions?

Comments?

…

Coffee?

Thanks for your attention!