***Fit covariates***

SPSS macros by Kirill Orlov

kior@comtv.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Подгонка переменных под матрицу коэффициентов.* Макросы изменяют значения переменных так, чтобы переменные имели силу связи в соответствии с указанной пользователем матрицей (корреляций, ковариаций или кросс-произведений). Опция страхования от гетероскедастичности позволяет добиться гомоскедастичных связей.

Для подгонки переменных к коварьированию друг с другом применяйте [!KO\_FITCOV](#_МАКРОС_!FITCOV:_ПОДГОНКА).

Для подгонки переменных к коварьированию с третьими переменными применяйте [!KO\_FITVAR](#_МАКРОС_!FITVAR:_ПОДГОНКА).

Оба макроса могут исправлять гетероскедастичность. Оба макроса относятся к данным как к непрерывным, ибо вольны изменять значения в дробные.

Для подгонки случайных данных сначала создайте эти случайные переменные. Например, так:

do repeat var= var1 to var10.

compute var= rv.normal(5,2).

end repeat.

execute.

# МАКРОС !KO\_FITCOV: ПОДГОНКА ПЕРЕМЕННЫХ ПОД КВАДРАТНО-СИММЕТРИЧЕСКУЮ МАТРИЦУ КОЭФФИЦИЕНТОВ

Version 1 (June 2003). Tested on SPSS 11, 11.5, 13, 14.

!KO\_fitcov vars= *v1 to v10* /\*Переменные для подгонки: список имен (можно ч-з to), ктр должны быть и среди имен столбцов matrix;

/\*либо ALL - взять все переменные файла и все столбцы матрицы, ктр тогда должно быть одинак кол-во, хотя имена не обязаны совпадать

/matrix= *'d:\exercise\corrmx.sav'* /\*Целевая матрица или матрица, содержащая ее в качестве подматрицы; внешний файл

/method= IPCOMP /\*Способ подправки ковариат: QUICK - неитерационно ч-з гл компоненты, IREGR - итерационно ч-з регрессию, IPCOMP - итерационно ч-з гл компоненты

/adjust= NONE /\*Для итерационных способов, дополнительные требования: NONE, CENTER (центровать непосредственно подгоняемую переменную - тж п/у),

/\*HSCED1 (страховать от веерной гетероскед-ти), HSCED2 (страховать от симметричной гетероскед-ти)

/iter= *20 2* /\*Для итерационных способов: Число итераций и, опционально, инкремент итераций

/msave= *'d:\exercise\newcorrmx.sav'* /\*Опционально: путь/имя внешнего файла для сохранения воспроизведенной матрицы.

Минимум должно быть задано VARS, MATRIX, METHOD, ITER

Макрос изменяет значения входящих переменных так, чтобы переменные показывали друг с другом силу связи, близкую или равную коэффициентам указанной пользователем матрицы, корреляционной, ковариационной или SSCP[[1]](#footnote-1). С помощью этого макроса можно «довести» существующие или случайные переменные до нужных величин коварьирования между ними. Кроме этого, наличествует опция исправления гетероскедастичности, т.е. доводка до гомоскедастичности[[2]](#footnote-2), в отношениях между переменными (это может иногда быть единственной целью применения макроса).

ПРИМЕР 1.

!KO\_fitcov vars= v1 v8 to v15 /matrix= 'd:\exercise\corrmx.sav' /method= IREGR /iter= 20 2.

* Переменные файла данных, v1, v8 и далее до v15 требуется подогнать к матрице коэффициентов “corrmx.sav” итеративным регрессионным методом с числом итераций 20 и инкрементом 2.

# *Процедура*

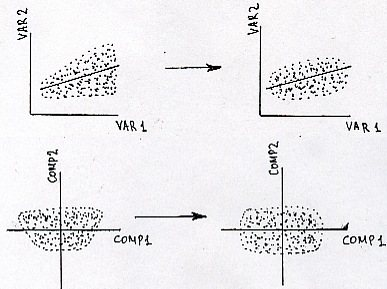
Подгонка переменных (ковариат) идет по итеративному алгоритму (придуманному автором макроса), заключающемуся в повторительном прибавлении к значениям данной ковариаты поправки, являющейся взвешенной средней тех поправок, которые требуют другие ковариаты для обеспечения нужных «им» коэффициентов связи с нашей ковариатой. Весами служат величины значений тех ковариат, благодаря чему можно постепенно удовлетворить «запросы» каждой из них. Этим достигается, что данная ковариата в итоге показывает связи с другими ковариатами, сколь угодно мало отличные от целевых (заданных) коэффициентов. Введение в формулу поправки новых членов и новых степеней для весов позволяет итеративно удовлетворить некоторым дополнительным требованиям, например гомоскедастичности (формулы см. ниже).

Итеративный алгоритм воплощен в макросе в двух «методах». Есть и третий метод, который не итеративен.

* *Регрессионный метод*. Вторая в списке входящая переменная берется как зависимая (предиктант), а 1-я как предсказывающая (предиктор, или регрессор). Должный между ними линейно-регрессионный коэффициент узнается из целевой матрицы и вычисляются предсказанные этой моделью значения, которые тут же вычитанием из предиктанта дают исходные регрессионные остатки. Эти остатки подгоняются итеративно к тому, чтобы иметь нулевое коварьирование с предиктантом (таким образом, целевыми *непосредственно* выступают нулевые коэффициенты). После подгонки и обеспечения должной их дисперсии остатки обратно складываются с предсказанными значениями в исправленную 2-ю переменную. После того, 3-я переменная берется за предиктант, а 1-я и 2-я за предикторы, и процедура повторяется теперь над 3-й переменной. И так далее: каждая переменная выступает зависимой от предшествующих, подгоняется (своими регрессионными остатками) и переходит в разряд предшествующих.
* *Главнокомпонентный метод.* Из целевой матрицы узнаем собственные числа и собственные векторы. Последние, перемножаясь с входящими переменными, дают заготовки главных компонент. Эти, начиная со 2-й, последовательно подгоняются итеративно (так же, как остатки подгоняются в регрессионном методе) к тому чтобы иметь нулевое коварьирование с предшествующими компонентами, – поскольку они должны стать взаимоортогональны. Заодно пересчитывается масштаб компонент на собственные числа целевой матрицы. По окончании подгонки компонент они обратно вращаются в переменные.
* *Неитеративный (быстрый) главнокомпонентный метод.* Входящие переменные вращаются в свои главные компоненты. Из целевой матрицы узнаем собственные числа и собственные векторы. Компоненты пересчитываются своим масштабом на целевые собственные числа, после чего компоненты обратно вращаются в переменные. Этот метод, заключающийся в пропорциональном сжатии/распускании облака данных вдоль компонент, скор и дает всегда полное достижение целевых коэффициентов ковариатами, но не обладает дополнительными возможностями.

Дополнительные возможности

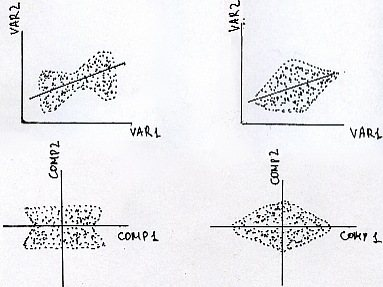
* *Центрация*. Во время подгонки переменной регрессионных остатков или компоненты накладывается требование достижения нулевой (близкой к 0) средней у этой, непосредственно подгоняемой переменной. Такое условие обычно весьма желательно: регрессионные остатки в норме должны иметь нулевую среднюю; а главные компоненты данных, за исключением 1-й компоненты, почти всегда имеют близкие или равные нулю средние. Опция центрации важна не как самоцель: при ней обыкновенно подгонка под коэффициенты успешнее.
* *Страхование от веерной гетероскедастичности*. Веерная гетероскедастичность довольно часто встречается в данных и делает ненадежной оценку регрессионных моделей. На рисунке, в верхнем ряду показана веерная гетероскедастичность зависимой переменной VAR2 вдоль предиктора VAR1, и то как она может быть исправлена в гомоскедастичность макросом: регрессионным методом с подключением опции «страховать от веерной гетероскедастичности». В этом случае макрос добивается (при достаточном числе итераций) гомоскедастичности текущей подгоняемой переменной в ее отношениях со всеми своими предикторами, т.е. предшествующими (уже подогнанными) переменными. Именно, *если подгоняемая сейчас своими остатками переменная – i-я, то ее изменчивость, «толщина» облака, - становится симметрична относительно центра всех i-1 предшествующих переменных.* Заметьте, что гомоскедастичность переменной *не гарантируется* в ее отношениях с последующими переменными, т.к. подгонка переменных при регрессионном методе идет последовательно.



Во втором ряду рисунка показано исправление веерной гетероскедастичности главнокомпонентным методом с опцией «страховать от веерной гетероскедастичности». В отличие от регрессионного метода, тут макрос добивается гомоскедастичности не самой подгоняемой сейчас компоненты, а всех предшествующих (уже подогнанных) компонент вдоль нее. Именно, *если подгоняемая сейчас компонента – i-я, то облако становится симметрично относительно ее центра в пространстве первых i компонент*. Так, на рисунке неравномерность изменчивости 1-й компоненты COMP1 вдоль 2-й компоненты COMP2, подгоняемой сейчас, заменена на равномерность. Гомоскедастичность последующих компонент (COMP3 и т.д.) не гарантируется вдоль подгоняемой COMP2.

На рисунке выше исправленные облака показаны «правильными». Однако в действительности страхование от веерной гетероскедастичности не гарантирует именно такой правильности, поскольку не препятствует случайному возникновению либо сохранению в данных Х-образной либо ромбической гетероскедастичности. Поэтому в макросе есть и более радикальная опция борьбы с гетероскедастичностью – страхование от симметричной гетероскедастичности.

* *Страхование от симметричной гетероскедастичности.* На рисунке ниже изображены симметричные виды гетероскедастичности – Х-образная (слева) и ромбовидная (справа). Опция «страхование от симметричной гетероскедастичности» подгоняет такие облака данных к гомоскедастичным, показанным на предыдущем рисунке справа. Веерная гетероскедастичность также исправляется этой опцией[[3]](#footnote-3). В сочетании с регрессионным методом опция делает так, что *если подгоняемая сейчас своими остатками переменная – i-я, то ее изменчивость, «толщина» облака, - становится равномерна вдоль всех i-1 предшествующих переменных.*



В сочетании с главнокомпонентным методом данная опция обладает способностью ликвидировать не только гетероскедастичность вдоль подгоняемой компоненты, но и гетероскедастичность самой подгоняемой компоненты вдоль предшествующих (т.е. исправляет двойную X-образность). Именно, *если подгоняемая сейчас компонента – i-я, то облако становится «равноутолщено» вдоль нее и вдоль всех предшествующих компонент.* Это означает, что после подгонки последней компоненты облако по всем направлениям многомерного пространства будет иметь «правильную», приближенную к эллипсоидной, форму.

Свойства методов

* *Регрессионный метод.* Единственный из 3-х методов, который не изменяет нелинейным образом первую переменную из списка входящих переменных. Он единственный позволяет вводить прямоугольные матрицы коэффициентов (см. макрос !KO\_FITVAR ниже). Достигнутая гомоскедастичность, если эта опция заказана, носит последовательный характер: каждая переменная гомоскедастична в отношениях с предшествующими в списке, но не обязательно гомоскедастична в отношениях с последующими в списке. С увеличением числа переменных подгонка последних в списке может быть заметно хуже, чем первых; это можно компенсировать инкрементом количества итераций. Если последующие планы исследователя связаны с регрессионным анализом, этот метод может оказаться предпочтительнее других (зависимую переменную лучше поместить в конце списка).
* *Главнокомпонентный метод.* При том же числе итераций как правило дает несколько лучшее достижение целевых коэффициентов, чем регрессионный метод (эта разница практически малозначима). Значения входящих переменных изменяет обычно меньше, чем два других метода, т.е. он наиболее «щадящий» по отношению к данным (что практически может быть важно). Ухудшение подгонки каждой следующей ковариаты малозаметно, т.к. подгоняются не ковариаты, а компоненты, которые потом поворачиваются в ковариаты, «размазывая» тем самым по ним недоподогнанность. Если целью является гомоскедастичность в парных отношениях между всеми переменными, то этот метод предпочтительнее регрессионного. Он также, очевидно, лучше, если данные предназначаются в дальнейшем для факторного анализа.
* *Неитеративный (быстрый) главнокомпонентный метод.* Этот метод, в отличие от итеративных, всегда дает точное достижение целевых коэффициентов. Входящие и итоговые ковариаты пропорционально связаны: каждая входящая переменная есть линейный многочлен всех итоговых, и каждая итоговая есть линейный многочлен всех входящих. Из трех методов, этот изменяет значения переменных обычно сильнее всех и, кроме того, наиболее неравномерно: большие по абсолютной величине значения изменяет относительно сильнее, чем это делают два других метода[[4]](#footnote-4).

***Результаты***

Макрос выводит в новый рабочий файл подогнанные ковариаты. Если в целевой матрице CORR или COV были ряды дополнительных статистик, ковариаты обладают ими. Если таких рядов не было, ковариаты выходят стандартизованными (при CORR матрице) или центрованными (при COV матрице).

При итеративных методах распечатывается информация о степени их подогнанности под целевую матрицу коэффициентов, т.е. насколько мало различаются целевая матрица и матрица, воспроизведенная итоговыми переменными. Этот показатель – «общая недостигнутость» - есть средняя по матрице абсолютная разница между соответствующими внедиагональными элементами целевой и воспроизведенной матриц. Для того чтобы показатель был стандартен независимо от типа целевой матрицы (корреляционная, ковариационная и т.д.), обе сравниваемые матрицы перед вычислением показателя переводятся в корреляционные матрицы.

Кроме недостигнутости, усредненной по всей матрице, распечатываются два постолбцовых ее показателя: усредненная недостигнутость в отношениях данной переменной со всеми предшествующими (эта информация имеет значение при регрессионном методе, показывая падение качества подгонки и тем самым помогая подобрать инкремент) и максимальные в столбцах недостигнутости – наибольшая абсолютная разница между двумя матрицами в данном их столбце. Диагональные элементы целевой и воспроизведенной матриц всегда полностью совпадают, и они не участвуют в вычислениях недостигнутости.

Макрос не выдает сведений о степени достижения гомоскедастичности. Интересующийся исследователь должен сам изучить ковариаты на этот предмет. Гетеро/гомоскедастичность бывает видна на графиках, сведения о ней можно получить из регрессионного анализа.

# *Требования к матрице*

Матрица целевых (требующих достичь их) коэффициентов должна быть квадратной симметрической. Обязательно присутствие столбцов ROWTYPE\_ (1-й) и VARNAME\_ (2-й). Других столбцов, кроме столбцов тела матрицы и этих, должно не быть. Имена переменных в рядах VARNAME\_ должны быть написаны *заглавными буквами* (в SPSS ранее 12-й версии) или должны *совпадать регистром* с именами переменных в рабочем файле данных (SPSS 12 и выше). Имена – не длиннее 8 букв.

Тип матрицы указывается значениями столбца ROWTYPE\_, которые должны быть написаны заглавными буквами. Следующие типы матриц допустимы:

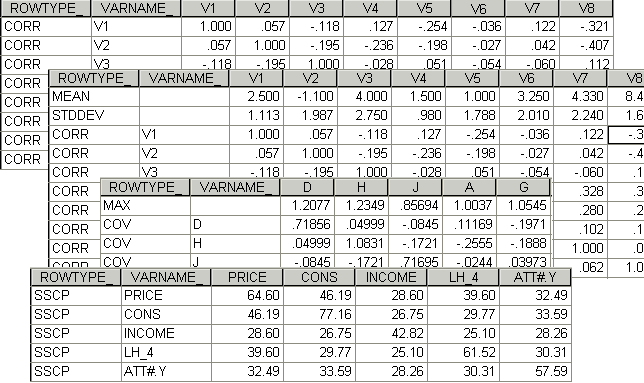
CORR корреляционная (на диагонали – единицы)

COV ковариационная (на диагонали – дисперсии)

SSCP сумм кросс-произведений (на диагонали – суммы квадратов)

MSCP средних кросс-произведений (на диагонали – средние квадраты), это SSCP матрица, деленная на размер выборки N

Матрицы CORR и COV могут, по желанию пользователя, иметь над собой дополнительные ряды статистик, к значениям которых требуется привести выходящие переменные. Тип статистики указывается заглавными буквами в ROWTYPE\_. Над COV может быть *один* ряд, любая из следующих статистик: MIN (наименьшее значение), MAX (наибольшее значение), MEAN (средняя арифметическая). Над CORR может быть *два* (не один) ряда: любое сочетание из следующих четырех статистик: MIN, MAX, MEAN, STDDEV (или SD; стандартное отклонение), причем из двух статистик *более ранняя в этом списке должна занимать 1-й ряд*. Примеры матриц:



Сверху вниз:

* Корреляционная матрица
* Корреляционная матрица с рядами дополнительных статистик: средняя и ст. отклонение
* Ковариационная матрица с рядом дополнительной статистики: наибольшее значение
* Матрица сумм кросс-произведений.

# *Подкоманды*

## MATRIX

Целевая матрица, внешний sav-файл – путь/имя в апострофах или кавычках. Пропущенные значения недопустимы в матрице.

## VARS

Переменные рабочего файла данных, которые требуется подогнать под коэффициенты целевой матрицы. Это могут быть собранные в исследовании или созданные случайно переменные. Их можно задать двояко:

*Список имен* - подгонять эти переменные (список можно писать поименно и/или ч-з to). Все эти переменные должны фигурировать среди имен столбцов/рядов целевой матрицы (в ней могут быть и другие имена). Порядок переменных в списке переменных или в файле не обязан совпадать с их порядком в матрице. В SPSS 12 и выше версий имена переменных в файле должны совпадать регистром букв с одноименными в целевой матрице.

ALL - подгонять все переменные рабочего файла. Их число должно равняться числу столбцов/рядов целевой матрицы, но имена там и там могут различаться: соответствующими будут считаться переменные и столбцы/ряды с одинаковым порядковым номером.

Если во входящих переменных есть пропущенные значения, макрос сразу заменит их на валидное значение 0. В ходе подгонки эти значения могут стать ненулевыми. Если замена на нуль исследователя не устраивает, пусть избавится от пропущенных данных заранее.

## METHOD

Укажите метод подгонки ковариат:

IREGR - Итеративный регрессионный. Непосредственно подгоняемые переменные это регрессионные остатки.

IPCOMP - Итеративный главнокомпонентный. Непосредственно подгоняемые переменные это главные компоненты.

QUICK - Быстрый главнокомпонентный. Итеративной подгонки нет.

## ADJUST

Укажите дополнительные требования подгонки (эта подкоманда не действует при METHOD=QUICK):

NONE - Никаких. Только достичь целевых коэффициентов.

CENTER - (тж. по умолчанию/незаданию подкоманды) Добиться нулевой средней у непосредственно подгоняемой переменной (при METHOD=IREGR это остатки, при METHOD=IPCOMP это 2-я и дальнейшие компоненты).

HSCED1 - Помимо требования CENTER[[5]](#footnote-5), добиться симметрии (в мере квадратов отклонений) облака рассеяния. Это страхование от веерной гетероскедастичности.

HSCED2 - Помимо требования CENTER, добиться равноутолщенности (в мере квадратов отклонений) облака рассеяния. Это страхование от симметричной гетероскедастичности.

## ITER

В этой подкоманде, которая не действует при METHOD=QUICK, задайте одно или два целых числа: начальное число итераций и, опционально, после него инкремент, на который должно увеличиваться число итераций при переходе к следующей подгоняемой ковариате/компоненте. Инкремент позволяет в определенной мере компенсировать ухудшение подгонки каждой очередной переменной/компоненты, связанное с «накоплением» недоподогнанности. Если переменных много, рекомендуется выше и инкремент. Инкремент имеет значение в основном при регрессионном методе.

Как начальное число итераций, так и инкремент имеет смысл делать больше при опциях достижения гомоскедастичности, т.к. такое достижение является дополнительным бременем, и если задать слишком мало итераций, ни коэффициенты, ни гомоскедастичность не будут достигнуты в удовлетворяющей пользователя мере. Удовлетворительное число итераций или инкремент нетрудно подобрать, глядя на распечатываемую макросом статистику «недостигнутости». Разумеется, от ITER и от количества переменных прямо зависит время исполнения макроса.

## MSAVE

Опционально: вы можете сохранить воспроизводимую ковариатами матрицу как внешний sav-файл. Для этого укажите путь/имя файла.

### Особые режимы

Макрос не рассчитан на расщепленное состояние файла. Взвешенность игнорируется макросом.

# *Некоторые вопросы*

*Как исправить гетероскедастичность, не меняя наблюдаемых сил связи.* Надо просто ввести целевую матрицу, в которой коэффициенты равны тем, что наблюдаются между переменными. Задействовать HSCED1 или HSCED2.

*Если подгонять случайные переменные, то какого распределения они должны быть?* Если нужно, чтобы выходящие ковариаты имели приблизительно нормальное распределение, нужно и входящие переменные брать из нормального или хотя бы симметричного распределения. Это особенно важно для переменной, которую вы хотите в дальнейшем рассматривать как зависимую от остальных ковариат в регрессионном анализе.

*Выдается сообщение, что целевая матрица не грамова.* Любая реальная (полученная из данных) ковариационная или кросс-продуктная матрица всегда является «положительно-(полу)определенной», или «матрицей Грама». Если ваша целевая матрица взята из головы или набита с ошибками, она может не быть такой. Это означает, что при данных диагональных величинах некоторые внедиагональные величины нереалистичны. Вы должны исправить матрицу.

*Выдается сообщение, что целевая матрица вырождена.* Это значит что некоторые ряды/столбцы матрицы однозначно предсказываются другими рядами/столбцами. Макрос не может работать с такой матрицей.

***Формулы подгонки***

Непосредственно подгоняемая переменная X (в регрессионном методе это регрессионные остатки, а в главнокомпонентном методе это заготовки главных компонент) на каждой итерации подправляется по формуле.

i наблюдение в файле (всего n наблюдений).

Vj каждая уже подогнанная раньше чем X переменная (их m).

Сj разница между наблюдаемой в начале итерации суммой кросс-произведений (мерой корреляции) X с Vj и таковой целевой. Т.к. согласно алгоритму все непосредственно целевые коэффициенты = 0, Cj это сама наблюдаемая сейчас сумма кросс-произведений X с Vj.

При ADJUST=NONE:



При ADJUST=CENTER. Перед итерацией X центруется (средняя приводится к 0).:



При ADJUST=HSCED1 или HSCED2. Перед итерацией X центруется.:



Si знак наблюдаемого значения Xi (+ или –). Этот множитель опускается при сочетании IPCOMP + HSCED1.

Wj центрованная переменная Vj – при сочетании IREGR + HSCED1; abs(Vj) – при сочетании IPCOMP + HSCED1; центрованная abs(центрованная Vj) – при сочетании IREGR + HSCED2; центрованная abs(Vj) – при сочетании IPCOMP + HSCED2.

Kj наблюдаемая в начале итерации сумма кросс-произведений X с Wj – при сочетании IPCOMP + HSCED1; abs(X) с Wj – при остальных сочетаниях.

На выходе из каждой итерации сумма квадратов X-переменной приводится к должной: Xi = Xi \* Sqrt( SSнужная / SSнаблюдаемая).

# МАКРОС !KO\_FITVAR: ПОДГОНКА ПЕРЕМЕННЫХ ПОД ПРЯМОУГОЛЬНУЮ МАТРИЦУ КОЭФФИЦИЕНТОВ

Version 1 (Nov 2005). Tested on SPSS 11, 11.5, 13, 14.

!KO\_fitvar cvars= *v1 to v5* /\*Имена переменных в раб файле, которые подгонять; эти имена должны быть среди имен столбцов matrix; можно писать ч-з to

/rvars= *w1 w4 w10* /\*Имена переменных в раб файле, к которым подгонять; эти имена должны быть среди имен рядов matrix; можно писать ч-з to

/matrix= *'d:\exercise\covmx.sav'* /\*Целевая прямоугольная матрица или матрица, содержащая ее в качестве подматрицы; внешний файл

/adjust= HSCED1 /\*Дополнительные требования в подгонке: NONE, CENTER (центровать непосредственно подгоняемую переменную - тж п/у),

/\*HSCED1 (страховать от веерной гетероскед-ти), HSCED2 (страховать от симметричной гетероскед-ти)

/iter= *20*  /\*Число итераций: целое положительное число

/msave= /\*Опционально: путь/имя внешнего файла для сохранения воспроизведенной матрицы.

Минимум должно быть задано CVARS, RVARS, MATRIX, ITER

Данный макрос, модификация !KO\_FITCOV, как и он, трансформирует переменные для достижения нужных значений коэффициентов связи (корреляции, ковариации или суммарные кросс-произведения). В отличие от !KO\_FITCOV, он подгоняет переменные к коварьированию не друг с другом, а с третьими переменными, которые сами никак не изменяются макросом. Входящими служат (1) одна или более подгоняемых переменных, (2) одна или более переменных, к которым подгоняются те, и (3) прямоугольная целевая матрица коэффициентов. Используется тот же итеративный алгоритм, что в !KO\_FITCOV, но только регрессионный метод возможен тут. Как и в !KO\_FITCOV, есть опция достижения гомоскедастичности (это может быть и единственной целью применения макроса).

ПРИМЕР 1.

!KO\_fitvar cvars= v1 v8 to v15 /rvars= w11 w19 /matrix= 'd:\exercise\corrmx.sav' /iter= 20.

* Переменные файла данных, v1, v8 и далее до v15 требуется подогнать к корреляциям с переменными w11 и w19, указанным в матрице коэффициентов “corrmx.sav”. Число итераций 20.

***Результаты***

Макрос выводит в новый рабочий файл подогнанные переменные. Если в целевой матрице CORR или COV были ряды дополнительных статистик, переменные обладают ими. Если таких рядов не было, переменные выходят стандартизованными (при CORR матрице) или центрованными (при COV матрице).

В окно результатов распечатывается информация о степени их подогнанности под целевую матрицу коэффициентов, т.е. насколько мало различаются целевая матрица и матрица, воспроизведенная итоговыми переменными. Этот показатель – «общая недостигнутость» - есть средняя по матрице абсолютная разница между соответствующими элементами целевой и воспроизведенной матриц. Для того чтобы показатель был стандартен, независимо от типа целевой матрицы (корреляционная, ковариационная и т.д.), обе сравниваемые матрицы перед вычислением показателя переводятся в корреляционные матрицы. Кроме общей недостигнутости, распечатываются максимальные в столбцах недостигнутости – наибольшая абсолютная разница между двумя матрицами в данном их столбце.

Макрос не выдает сведений о степени достижения гомоскедастичности. Интересующийся исследователь должен сам изучить связи на этот предмет. Гетеро/гомоскедастичность бывает видна на графиках, сведения о ней можно получить из регрессионного анализа.

# *Требования к матрице*

Матрица должна быть «асимметрической прямоугольной», т.е. ее столбцы и ряды задают разные списки переменных, а не один список. Переменные, которые требуется подогнать, должны соответствовать *столбцам* матрицы, а переменные, к связям с которыми их надо подогнать, - *рядам* матрицы. Имена тех и других – не длиннее 8 букв. Обязательно присутствие столбцов ROWTYPE\_ (1-й) и VARNAME\_ (2-й). Других столбцов, кроме столбцов тела матрицы и этих, должно не быть. Имена переменных в рядах VARNAME\_ должны быть написаны *заглавными* буквами (в SPSS ранее 12-й версии) или должны *совпадать регистром* с именами переменных в рабочем файле (SPSS 12 и выше версии). Имена переменных-столбцов матрицы должны *совпадать регистром* с именами переменных в рабочем файле (SPSS 12 и выше версии).

Тип матрицы указывается значениями столбца ROWTYPE\_, которые должны быть написаны заглавными буквами. Следующие типы матриц возможны:

CORR корреляционная

COV ковариационная

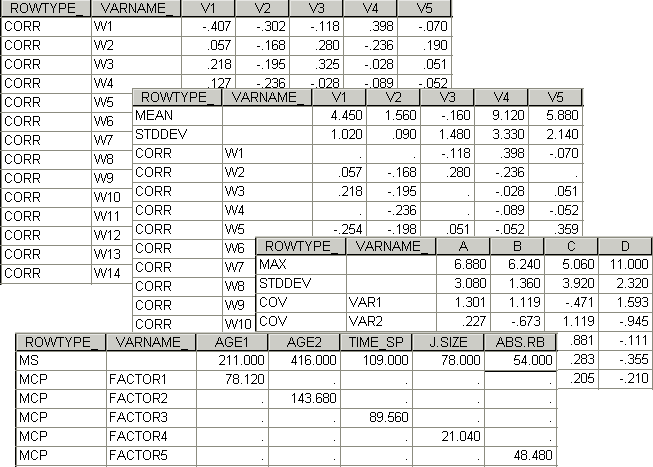
SCP сумм кросс-произведений

MCP средних кросс-произведений; это SCP, деленные на размер выборки N

Над телом матрицы располагается этот *обязательный* ряд, также обозначенный в ROWTYPE\_:

* Над SCP матрицей должен быть ряд SS (суммы квадратов)
* Над MCP матрицей должен быть ряд MS (средние квадраты)
* Над COV матрицей должен быть ряд STDDEV (или SD; стандартное отклонение)
* Над CORR матрицей нет обязательного ряда

Еще выше *могут* находиться, в случае матриц COV и CORR, дополнительные ряды статистик, к значениям которых требуется привести выходящие переменные. Для COV возможен *один* ряд, любая из следующих статистик: MIN (наименьшее значение), MAX (наибольшее значение), MEAN (средняя арифметическая). Над CORR может быть *два* (не один) ряда: любое сочетание из следующих четырех статистик: MIN, MAX, MEAN, STDDEV (или SD; стандартное отклонение), причем из двух статистик *более ранняя в этом списке должна занимать 1-й ряд*. Примеры матриц:



Сверху вниз:

* Матрица корреляций
* Матрица корреляций с дополнительными статистиками: средняя и ст. отклонение
* Матрица ковариаций с дополнительной статистикой: наибольшее значение; ряд STDDEV обязателен
* Матрица средних кросс-произведений; ряд MS обязателен

В теле целевой матрицы разрешены *пропущенные значения* (system- или user-missing). Пользователь, иначе говоря, может оставлять в теле матрицы пустые ячейки на месте коэффициентов, к которым не требуется подгонять переменные. Так, на рисунке выше 2-я сверху матрица требует подогнать переменную V1 к целевым корреляциям с переменными W2, W3, W5, но не требует подогнать ее к определенным корреляциям с W1 и W4. Корреляции V1 с этими последними исследователя (и макрос) не интересуют, - и они получатся такие, какие получатся. Другой пример: на нижней матрице указаны коэффициенты только на диагонали, только они и составляют целевую матрицу. Задача поэтому чисто двупеременная: создать нужную силу связи AGE1 с FACTOR1, AGE2 с FACTOR2, и т.д. Пропущенные значения *не разрешены* в обязательном ряду и в дополнительном ряду, которые лежат над матрицей. В матрице должно не быть значения: ровно *–999*, т.к. оно зарезервировано макросом для своих нужд.

# *Подкоманды*

## MATRIX

Целевая матрица, внешний sav-файл – путь/имя в апострофах или кавычках.

## CVARS, RVARS

CVARS это поименный или ч-з to список переменных рабочего файла данных, которые требуется подогнать под целевые коэффициенты. Все эти имена должны присутствовать среди имен столбцов целевой матрицы (которых может быть больше); порядок их в списке или в файле не обязан совпадать с порядком в матрице и не сказывается на результатах. Аналогично, RVARS это список переменных файла, к коварьированию с которыми нужно подгонять CVARS. Их имена должны присутствовать среди рядов целевой матрицы. В SPSS 12 и выше версий имена переменных в файле должны совпадать регистром букв с одноименными в целевой матрице. Переменные RVARS не изменяются макросом и не выдаются обратно на выход. Любой из двух списков может состоять даже из одной переменной.

Пропущенные значения в RVARS недопустимы, а в CVARS допустимы – они будут сразу заменены на валидный нуль. В ходе подгонки эти значения могут стать ненулевыми. Если замена на нуль исследователя не устраивает, пусть он избавится от пропущенных данных заранее.

## ADJUST

Укажите дополнительные требования подгонки:

NONE - Никаких. Только достичь целевых коэффициентов.

CENTER - (тж. по умолчанию/незаданию подкоманды) Добиться нулевой средней у подгоняемых регрессионных остатков.

HSCED1 - Помимо требования CENTER, добиться симметрии облака рассеяния. Это страхование от веерной гетероскедастичности.

HSCED2 - Помимо требования CENTER, добиться равноутолщенности облака рассеяния. Это страхование от симметричной гетероскедастичности.

## ITER

Задайте число подгоночных итераций каждой переменной.

## MSAVE

Опционально: вы можете сохранить воспроизводимую переменными матрицу как внешний sav-файл. Для этого укажите путь/имя файла. В матрице нет пустых ячеек. Коэффициенты, соответствующие пустым ячейкам целевой матрицы, это те, к которым переменные не подгонялись. Эти коэффициенты не участвуют в вычислении показателя недостигнутости.

### Особые режимы

Макрос не рассчитан на расщепленное состояние файла. Взвешенность игнорируется макросом.

1. Общее название класса коэффициентов, о которых идет речь – «скалярные произведения векторов-переменных» или «суммарные кросс-произведения». [↑](#footnote-ref-1)
2. Гетероскедастичность – нежелательное для статистика явление, когда вариативность признака непостоянна вдоль диапазона значений другого признака, с которым он связан. Гомоскедастичность – противоположное, обычно желательное явление – когда вариативность постоянна. [↑](#footnote-ref-2)
3. Если вы не ожидаете от данных симметричной гетероскедастичности или не заботитесь о ней, а преследуете только веерную, то лучше воспользоваться все же опцией «страхование от веерной гетероскедастичности» как более подходящей к этому случаю. [↑](#footnote-ref-3)
4. Известен еще неитеративный, тоже точный метод с использованием корня Холецкого целевой матрицы, однако, по моим наблюдениям, он еще сильнее изменяет значения входящих переменных, чем неитеративный главнокомпонентный способ, т.е. уступает ему, если стоит задача – *подправить* входящие переменные под целевую матрицу. [↑](#footnote-ref-4)
5. В условиях HSCED1 и HSCED2, в методе IPCOMP 1-я компонента тоже центруется, как и остальные [↑](#footnote-ref-5)