R101: A practical guide to making R your everyday statistical tool (PSY532)

Lecture 7

Requested and supplementary material

Programme

- Factor analysis
  - Logic of the analysis
  - Conducting the analysis: 11 steps
  - Two examples (one just in the code, and with an accompanying reading)
- Further analysis of interaction effects in ANOVA
  - Interpreting the interaction contrasts provided by the `summary` function for `lm`, `glm` and `bayesglm`
  - Exploratory analyses and significance tests in the `phia` package

Factor analysis

Logic of the analysis/ Key concepts

Readings: 1 (Field), 2 (Costello & Osborne); others specified in the slides

- Factor analysis identifies groups of observed variables that tend to hang together empirically. The variables are usually items in a questionnaire.
- Purposes:
  - understand the structure of a set of variables (e.g., intelligence)
  - development of new questionnaires: do the items have an expected factor structure? (e.g., Do the items of a new intelligence test capture the known distinction between “fluid” and “crystallised” intelligence?)
  - data set reduction, often involving the calculation of factor scores

  The factor analyses in this lecture have all of these purposes.
- Two main types:
  - Principal component analysis (not technically “factor analysis”)
  - Factor analysis (usually, with a principal axis factoring fitting procedure, but maximum likelihood available as well)

  Tend to produce similar results, but not in all cases. So analyses typically examine the results from both methods.
- Performed on a correlation or covariance matrix of the observed variables.

Key concepts: Factor loadings

- One of the main final "outputs" of the regression analysis
- Structure matrix: loadings expressed as the average correlation, across all participants *i*, of score on the observed variable and score on the factor
- Pattern matrix: loadings expressed as regression coefficients:

  Illusion of natural control$_i$ = 0.95 x Practice$_i$ + 0.75 x GoalieMoves$_i$ + 0.69 x Skill$_i$ … + 0.39 x Luck$_i$ + error$_i$

  Illusion of supernatural control$_i$ = -0.13 x Practice$_i$ +… 0.63 x LuckyPlay$_i$ + 0.59 x LuckyMoments$_i$+ 0.43 x Luck$_i$ + error$_i$
- Correlations and regression coefficients are the same thing when an orthogonal rotation is used (see "Rotation" concept).

Key concepts: Factor scores (Reading: 3)

- A person's score on a factor could be calculated as just the average of all the items loading on that factor, but this can be a suboptimal measure given that the items have different weights.
- There are various ways of calculating factor scores, the most common one being "regression", which takes correlations between factors and items into account.

Key concepts: Communality, and its role in distinguishing principal component analysis from principal axis factoring

- **Principal axis factoring (PAF)/ maximum likelihood** assumes that the covariation in the observed variables is due to the presence of one or more latent variables (factors) that exert **causal influence** on these observed variables. Under this assumption, each observed variable can have variance *not* caused by the latent variable. This unique variance can occur because of measurement error (random variance) or reliable individual differences not related to the latent variable. In mathematical terms, this translates into the assumption that, for each observed variable, the **communality (a measure of the variance shared with other observed variables) is not necessarily equal to one**. The initial communality is the squared multiple correlation of each variable with all the others (the R^2 if all the other variables are assumed to be predictors in a linear regression). See this webpage for the details.

- **Principal component analysis (PCA)** makes no assumptions about an underlying causal model. It is simply a variable reduction procedure aimed at identifying a relatively small number of components that account for most of the variance in a set of observed variables. **Communalities are assumed to equal one.**

Key concepts: Eigenvalues

- Eigenvalues are calculated for each uncovered factor/component during both PAF and PCA. PAF just uses a slightly different correlation matrix, where the initial communalities replace the '1s' along the diagonal. There are always as many factors/components as observed variables, and the first factor/component always accounts for the largest amount of total variance relative to the other factors.
- Eigenvectors consist of the "weights" (loadings) of each observed variable on the factor. Eigenvalues are a property of the eigenvectors and have a complex geometrical definition. The higher a factor's eigenvalue, however, the greater the amount of total variance it accounts for.
- Eigenvalues can be illustrated in a scree plot. This plot might show slightly different values, depending on whether it illustrates the PAF or PCA calculations. However, the general shape of a scree plot is always the same, illustrating, as per the first dot point, higher relative importance for the first factor.
- As will be discussed later in "Analysis steps", scree plots are used to decide whether to retain more than one factor.

Key concepts: Rotation

- Performed so that the loadings of the observed variables are maximised on the factor that they relate to most.
- Can be:
  - Orthogonal (varimax, quatrimax, BentlerT, geominT): factors are assumed to be independent (i.e., at right angles, as in the diagram)
  - Oblique (olimin, promax, simplimax, BentlerQ, geominQ): factors are assumed to be related
- Often, both are performed, with the "factor correlation matrix" then being checked to determine whether there are substantial correlations between factors that warrant discarding the orthogonal solution. Factors in human data tend to be correlated.
- Factor rotation methods matter: e.g., promax is recommended over oblimin for very large data sets.

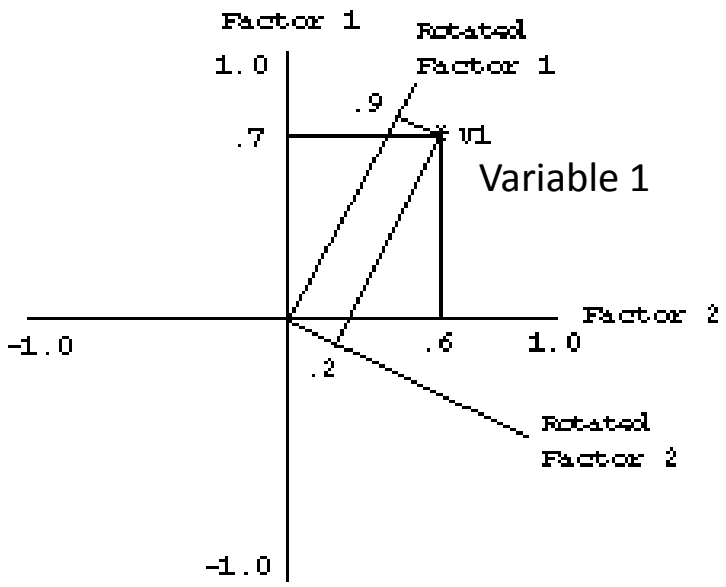

Image from:
http://www.ats.ucla.edu/stat/sas/library/factor_ut.htm

Key concepts: Sample size

- Often discussed in terms of the ratio of subjects to observed variables/questionnaire items:
  - In practice, researchers use factor analysis with ratios as low as 2:1, with most using 5:1 and 10:1. The "classic" recommendation is to have between 5 and 10 times the number of participants as items.
- However, recent simulation studies suggest that the above recommendation is too simplistic:
  - Reading 4 (MacCallum et al.) suggests, for example, that sample size matters when factor loadings are low or factors are not easily distinguishable from each other.
- The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy can be used to gauge the degree to which any obtained factors are likely to be "distinguishable". The KMO is the ratio of the squared correlation between variables to the squared partial correlation. The partial correlation for each pair of variables corresponds to the correlation between those variables after partialling out the influence of all of the other variables in the factor analysis. If the variables share distinct common factor(s), then the partial correlations should be small and the KMO should be close to 1 (rather than 0).

Analysis steps

Reading: 1 (Field)

1. Determine the factor structure you expect and some possible alternatives. A one-factor solution is the obvious comparison structure, but there might be others. For example, you might expect the questionnaire items you are using to load on two factors, as has been found in most previous studies. However, one previous study might have obtained a three-factor solution.
2. If considering using a maximum likelihood fitting procedure, check that responses on the items are normally distributed. We will use principal axis factoring here, because it is more common, and because responses in the first example are not normally distributed.
3. Remove two kinds of items:
   - Weak items: Those that do not correlate highly with many other items (i.e., for which $r = 0.3$ or less in many cases). Look at these items to see what's wrong with them.
   - Singular items: Those that correlate highly (> .8) with any other item, suggesting multicollinearity.

4. Run formal tests of factorability:
   - KMO (mentioned on previous slide)
   - Bartlett's test of sphercity (are all the items quite weak?): If no, the correlation matrix should be significantly different from an identity matrix.
5. Determine the number of factors to extract using a scree plot and parallel analysis. "Parallel analysis" runs a preliminary (unrotated) principal components analysis (PCA) or princial axis factoring analysis (PAF) on your data. It also does this for a random data set with the same number of observed values. The two sets of eigenvalues are then compared. If the eigenvalues from the random data are larger then the eigenvalues from the PCA, the components/factors are concluded to be random noise and a one-factor solution is recommended.
6. Run a PCA with orthogonal and oblique rotations.
7. Run PAF analyses with the same rotations and determine which results to treat as "final" based on your understanding of the debate between defenders of PCA and PAF.
8. If, in your chosen solution, there are a small number of items that have loadings on more than one factor because of some ambiguity in the wording, remove those items and re-run the analysis (the single one you have chosen).
9. Interpret the item loadings in your final solution. What factors/ components do the items suggest?
10. Compute factor scores if they are needed for a subsequent analysis.
11. Calculate Cronbach's alpha, a measure of split-half reliability (see reading), for your obtained factors/components.

Example 1: Illusions of natural and supernatural control in the SS data

Reading: 5 (Ejova et al.)

Is the following clustering of the items supported by exploratory factor analysis? Can we calculate factor scores to use as outcome measures instead of the item averages we have been using throughout the course?

When thinking about your wins/goals, to what extent would you use each of the following statements to describe how they came about? (0 to 10 for each statement)

'Natural' IoC

1. My skill in playing the game.
2. I got better with practice.
3. My knowledge of soccer.
4. I developed a logical strategy for playing.
5. Experience in playing computer games.
6. The players I chose.
7. The kick directions I chose.
8. I learned how to predict the movements of the goalkeeper.

'Supernatural' IoC

1. I took advantage of moments when my luck was good.
2. I've always been a lucky kind of person.
3. I deserved to win.
4. I knew how to make my luck turn good.
5. A certain lucky way of playing just seemed to work for me.

- The packages you'll need:
  - `psych`
  - `nFactors`
  - `GPArotation`
- Refer to the script for code pertaining to each of the steps
- Note regarding Step 1: Based on studies where people who gamble were interviewed or surveyed, we expect the statements comprising the questionnaire to form two groups, representing the illusions of natural and supernatural control. If the statements are vague, however, and people understood them in different ways, there might be only one factor.
- Note regarding Step 3: As Table C.1 in the Reading shows, there were many zero ratings for each statement (and many ratings of "10" for "It was all chance"). Thus, each item was screened for whether scores on it correlated with other items when only participants *without* zero ratings on the item were considered. The "Chance" item was excluded based on this screening, as shown in the script.
- Factor Analysis Example 2 is in the script for those who are interested in a similar analysis involving more items. The motivation for the analysis is in Reading 6.

More on interaction effects

Interpreting interaction contrasts provided by the `summary` function

Interactions between categorical variables in linear models

Let's consider the analysis for testing Hypothesis 1 in the SS dataset. You encountered it in Assignment 2. In the answer sheet, I suggested making prior beliefs a categorical variable and then testing its effect in interaction with the main predictor of interest – success slope type (descending vs. U-shaped vs. ascending, etc.) Remember that the illusion of natural control is the average of eight items, each of which have a maximum value of 10. Thus, our outcome measure also has a maximum value of 10.

```
#Make prior beliefs (PreDBC_IOC) a categorical variable:
SS$CatPreDBC_IOC <- cut(x = SS$PreDBC_IOC, breaks = 2, labels =
c("Low", "High"))

#Run ANOVA allowing for main effects and interaction
anova1 <- aov(PostNaturalIoC ~ SeqCond*CatPreDBC_IOC, data =
SS)
summary.lm(anova1)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**1.74330**	0.24602	7.086	8.57e-12	***
SeqCondU	**0.04836**	0.35113	0.138	0.8905	
SeqCondAsc	0.57641	0.35455	1.626	0.1050	
SeqCondFlat	0.06062	0.35635	0.170	0.8650	
CatPreDBC_IOCHigh	**0.99808**	0.42119	2.370	0.0184	*
SeqCondU:CatPreDBC_IOCHigh	**1.24151**	0.58829	2.110	0.0356	*
SeqCondAsc:CatPreDBC_IOCHigh	0.95073	0.60672	1.567	0.1181	
SeqCondFlat:CatPreDBC_IOCHigh	1.03891	0.58870	1.765	0.0785	.

Intercept/constant: Illusion of natural control (on a 10-point scale) when all predictors and interaction terms are equal to zero.

Example of main effect contrast 1: Number of units out of 10 by which the illusion of natural control is larger when the success slope is U-shaped rather than Descending.

Example of main effect contrast 2: Number of units out of 10 by which the illusion of control is larger when prior belief in gambling game controllability is high rather than low.

Example of a two-way interaction contrast: Number of units out of 10 by which the illusion of natural control is larger when the success slope is U-shaped and prior beliefs are high, as opposed to when the success slope is Descending and prior beliefs are low.

The second described main effect contrast (for prior beliefs) and the described interaction contrast are significant.

Interactions between categorical and numeric variables in linear models

Let's re-run the previous analysis, but with prior beliefs as a numeric (interval-scale) variable showing total score on the "illusion of control" scale of the *Drake Beliefs About Chance Inventory.*

```
anova2 <- aov(PostNaturalIoC ~ SeqCond*PreDBC_IOC, data = SS)
summary.lm(anova2)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**0.28750**	0.60806	0.473	0.63666	
SeqCondU	**-1.41429**	0.85977	-1.645	0.10094	
SeqCondAsc	-1.83316	0.96936	-1.891	0.05950	.
SeqCondFlat	-1.27483	0.86497	-1.474	0.14149	
PreDBC_IOC	**0.06482**	0.02086	3.108	0.00205	**
SeqCondU:PreDBC_IOC	0.07154	0.02961	2.416	0.01623	*
SeqCondAsc:PreDBC_IOC	**0.09298**	0.03286	2.830	0.00495	**
SeqCondFlat:PreDBC_IOC	0.06307	0.02948	2.139	0.03314	*

Intercept/constant: Illusion of natural control (on a 10-point scale) when all predictors and interaction terms are equal to zero.

Example of main effect contrast 1: Number of units out of 10 by which the illusion of natural control is larger when the success slope is U-shaped rather than Descending.

Example of main effect contrast 2: Number of units out of 10 by which the illusion of control is smaller when prior belief in gambling game controllability increases by one unit on the *Drake Beliefs About Chance* scale.

Example of a two-way interaction contrast: Number of units out of 10 by which the illusion of natural control is larger when the success slope is Ascending rather than Descending and prior beliefs increase by one unit.

The second described main effect contrast (for prior beliefs) and all two-way interaction contrasts are significant.

Interactions in generalised linear models

Let's consider the generalised linear modelling analysis from Assignment 4, where we examined reported average monthly gambling expenditure as a function of three predictor variables: whether one has ever played online (yes/no), whether one is a young male (yes/no), and income (low, average, high).

```
Sp1 <- Spending[(Spending$Spend < 10000),]
       #creating a new data frame with outliers removed
library(MASS)
glmnba <- glm.nb(Spend ~ Online*YoungMale*IncomeCat,
       data = Sp1)
summary(glmnba)
```

	Estimate	Std. Err	z value	Pr(>|z|)
(Intercept)	**4.7612**	0.3555	13.392	< 2e-16 ***
OnlineYes	**0.3734**	0.4898	0.762	0.445828
YoungMaleYes	0.2494	0.4315	0.578	0.563225
IncomeCatMid (salary 10-20 000)	1.3943	0.4190	3.328	0.000876 ***
IncomeCatHigh (salary over 20 000)	1.2857	0.4440	2.896	0.003780 **
OnlineYes:YoungMaleYes	1.1307	0.6020	1.878	0.060369 .
OnlineYes:IncomeCatMid (salary 10-20 000)	-0.4452	0.6037	-0.737	0.460897
OnlineYes:IncomeCatHigh (salary > 20 000)	**-0.3361**	0.6605	-0.509	0.610882
YoungMaleYes:IncomeCatMid (salary 10-20 000)	-0.4964	0.5430	-0.914	0.360661
YoungMaleYes:IncomeCatHigh (salary > 20 000)	-0.4253	0.5865	-0.725	0.468320
OnlineYes:YoungMaleYes:IncomeCatMid	**-0.5555**	0.7733	-0.718	0.472575
OnlineYes:YoungMaleYes:IncomeCatHigh	-0.3544	0.8561	-0.414	0.678868

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Intercept/constant: Expenditure (in log counts) when all predictors and interaction terms are equal to zero. Log counts means that, if you raise *e* (approx. 2.718) to the power of the estimate, you get the coefficient in the original units (Czech Crowns): 1.56.

Example of main effect: *e*^0.373 is the number of Crowns by which the expenditure is larger if one is a young male rather than a woman or older male.

Example of a two-way interaction effect: *e*^0.336 is the number of Crowns by which expenditure is smaller if one has never played online and earns a low income, compared to when one has had online experience and earns a high income.

Example of a three-way interaction effect: *e*^0.556 is the number of Crowns by which expenditure is smaller if one has never played online, is not a young male, and has low income, compared to when one has played online, is a young male and has 'mid' income.

None of the interaction effects are significant in this case.

Significance tests in the `phia` package

Reading: 6

- The `phia` package can be used for graphing interaction effects and testing the significance of the group differences apparent in the plots.
- These mean differences might not be automatically tested in `summary`. In `phia`, it is possible to indicate a wide variety of group differences.
- The script presents examples for generalized linear models encountered in Lecture 4 and Assignment 4 (as well as Assignment 2). The first interaction we examine is the one interpreted in the two previous slides.
- Reading 6 provides excellent example code, with a special section on generalized linear models.
- The next two slides provide clarification where Reading 6 is a little technical.

A note on the “fixed” and “pairwise” arguments

- In the command `testInteractions(glmnba, fixed="YoungMale", across="Online")`, using the argument “fixed” means looking at the simple effects of Online (yes/no) at the two levels of YoungMale (yes/no) separately. For linear models, *t*-tests or *F*-tests are conducted, depending on whether the “across” variable has two or more groups. The *F*-test is used in the latter situation. How many *F*-tests or *t*-tests are conducted depends on the number of groups in the “fixed” variable (two in our example, leading to two *t*-tests or *F*-tests). For generalized linear models, such as the one in our example, *t*-tests and *F*-tests are replaced by Wald chi-square tests. A correction (Holm) is applied to adjust for the effects of multiple testing.
- In the command `testInteractions(glmnba, pairwise = "YoungMale", across="Online")`, using the argument “pairwise” means testing the significance of the difference of differences. That is, we use the *t*-distribution or *F*-distribution (for linear models) or a *chi-squared* distribution (for generalized linear models) as the sampling distribution in testing whether the following difference is significantly different from zero:

  $[m(0, 1) - m(0, 0)] - [m(1, 1) - m(1, 0)]$

  Here, the first variable in the brackets is the variable specified for pairwise (YoungMale), while the second variable is the variable specified for across (Online). 0 and 1 refer to “yes” and “no”, respectively. So, here, we’re testing whether [NotYoungMaleOnline - NotYoungMaleNotOnline] - [YoungMaleOnline - YoungMaleNotOnline] is significantly different from zero.

Creating a `custom.contr` matrix

Creating a custom contrasts matrix in `phia` means creating different meanings for the first and/or second variables. In our examples, we create different meanings just for the first variable, SeqCond. Since we are not specifying "pairwise" and "across" variables any longer, it's more useful to think of the first variable as the one labelled in the rows of the output, and of the second variable as the one labelled in the columns or not seen.

So, in the first row of the output on line 343 of the script, the row variable is "Asc vs. all" with 0 meaning "Ascending" and 1 meaning "All". In the second row, the row variable is Asc vs. Desc where 0 means Ascending and 1 means "All". The column variable in both rows is question wording (CaptionType). Zero on this variable is the reference level, which can be found by writing: levels(SS$PostHowManySingleCaptionType). The first level – the reference level – is "goals", so this is the wording corresponding to 0 on this variable.

$[m(0, 1) - m(0, 0)] - [m(1, 1) - m(1, 0)]$

Readings

All available as pdfs in Study Materials/Lecture 7.

1. Field, A., Miles, J., & Field, Z. (2012). *Discovering Statistics Using R.* Sage: UK. Chapter 17. Exploratory Factor Analysis.
2. Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10.
3. DiStefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation*, 14.
4. MacCallum, R. C., Widaman, K. F., Zhang, S. & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*, 84-99.
5. Ejova, A., Navarro, D. J., & Delfabbro, P. H. (2013). Success-slope effects on the illusion of control and on remembered success-frequency. *Judgment and Decision Making, 8,* 498–511.
6. Ejova, A., Delfabbro, P. H., & Navarro, D. J. (2013). Erroneous gambling-related beliefs as illusions of primary and secondary control: A confirmatory factor analysis, *Journal of Gambling Studies.* Available only as an electronic pre-publication.
7. De Rosario Martinez (2013). Analysing interactions of fitted models. *R* supplementary documentation.