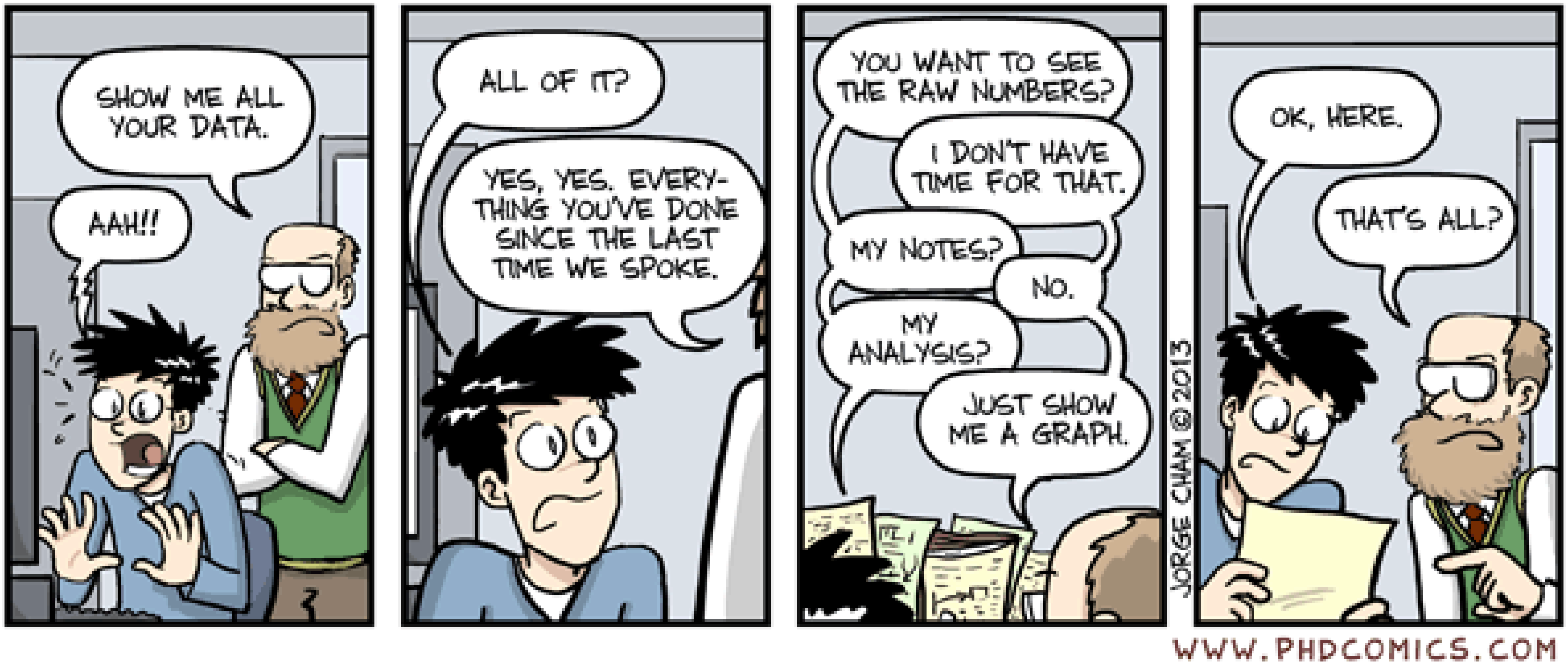
Workshop:
The basics of Factor Analysis and PCA



February 28, 2020

Agenda

1) What is Factor Analysis and PCA?

2) How to run it on Stata or R? (More on this next week)

3) Interpreting the outputs.

Agenda

1) What is Factor Analysis and PCA?
2) How to run it on Stata or R? (More on this next week)
3) Interpreting the outputs.

Introduction

Measuring latent constructs.

ST094

How much do you disagree or agree with the statements about yourself below?

(Please select one response in each row.)

	Strongly Disagree	Disagree	Agree	Strongly Agree
I generally have fun when I am learning science topics.	○	○	○	○
I like reading about science.	○	○	○	○
I am happy working on science topics.	○	○	○	○
I enjoy acquiring new knowledge in science.	○	○	○	○
I am interested in learning about science.	○	○	○	○

Science aptitude?

Introduction

Measuring latent constructs.

How to use these data?

1. Average?
2. Weighted average?
3. Other strategies?

Factor Analysis or Principal Component Analysis!

😱

The intuition

Measuring latent constructs – UNOBSERVED!

Small step back. What does regression do?

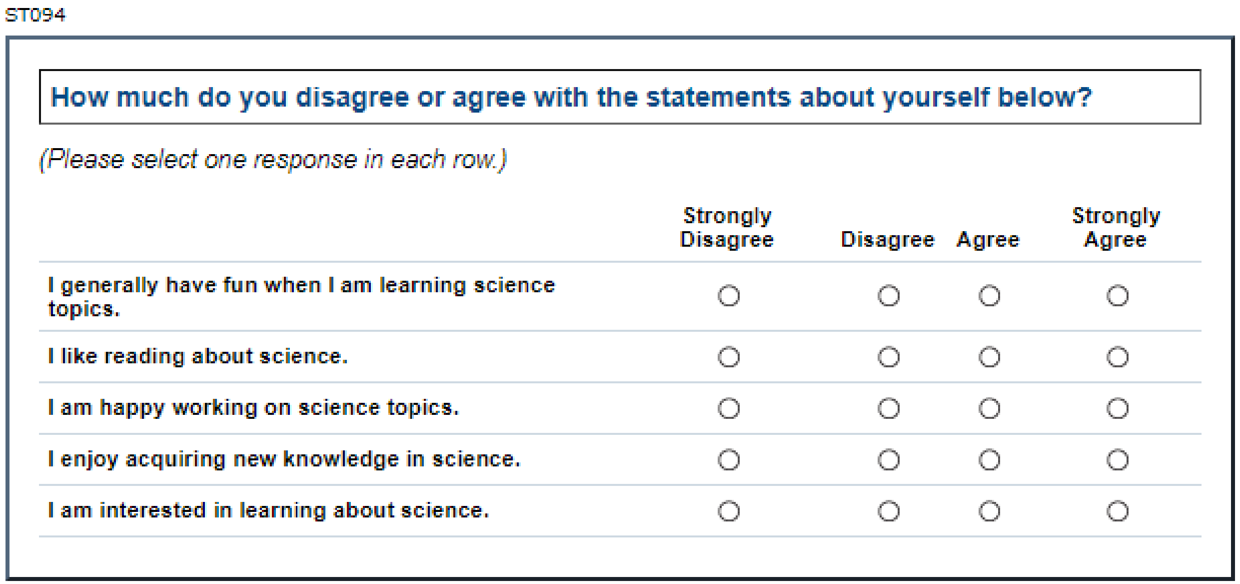
Explains the variation in the dependent variable.

$$A_1 = \beta_0 + \beta_1 Var_1 + \beta_2 Var_2 + \beta_3 Var_3 + \cdots + \beta_m Var_m + \epsilon_1$$

The problem is that we have more than one A!

$$A_1 = \mu_1 + l_{11} f_1 + l_{12} f_2 + l_{13} f_3 + \cdots + l_{1m} f_m$$
$$A_2 = \mu_2 + l_{21} f_1 + l_{22} f_2 + l_{23} f_3 + \cdots + l_{2m} f_m$$
$$A_3 = \mu_3 + l_{31} f_1 + l_{32} f_2 + l_{33} f_3 + \cdots + l_{3m} f_m$$
$$A_4 = \mu_4 + l_{41} f_1 + l_{42} f_2 + l_{43} f_3 + \cdots + l_{4m} f_m$$
$$A_5 = \mu_5 + l_{51} f_1 + l_{52} f_2 + l_{53} f_3 + \cdots + l_{5m} f_m$$

$$A_i = Answer\ to\ Q_i$$
$$f_i = Factor\ i$$
$$l_{qi} = Factor\ loading\ of\ f_i\ for\ question\ Q_i$$


ST094

How much do you disagree or agree with the statements about yourself below?

(Please select one response in each row.)

	Strongly Disagree	Disagree	Agree	Strongly Agree
I generally have fun when I am learning science topics.	○	○	○	○
I like reading about science.	○	○	○	○
I am happy working on science topics.	○	○	○	○
I enjoy acquiring new knowledge in science.	○	○	○	○
I am interested in learning about science.	○	○	○	○

The intuition – part 2

Each factor f can explain the variation in each A by the estimated loading l_q. As we have more than one A there are more than one estimated f and l.

$$
\begin{aligned}
A_1 &= \mu_1 + l_{11}f_1 + l_{12}f_2 + l_{13}f_3 + \cdots + l_{1m}f_m \\
A_2 &= \mu_2 + l_{21}f_1 + l_{22}f_2 + l_{23}f_3 + \cdots + l_{2m}f_m \\
A_3 &= \mu_3 + l_{31}f_1 + l_{32}f_2 + l_{33}f_3 + \cdots + l_{3m}f_m \\
A_4 &= \mu_4 + l_{41}f_1 + l_{42}f_2 + l_{43}f_3 + \cdots + l_{4m}f_m \\
A_5 &= \mu_5 + l_{51}f_1 + l_{52}f_2 + l_{53}f_3 + \cdots + l_{5m}f_m
\end{aligned}
$$

$$
\mathbf{f} = \begin{bmatrix} \mathrm{f}_1 \\ \mathrm{f}_2 \\ \mathrm{f}_3 \\ \vdots \\ \mathrm{f}_5 \end{bmatrix} \; and \; \boldsymbol{l} = \begin{bmatrix} l_{11} l_{12} \ldots l_{15} \\ l_{21} l_{22} \ldots l_{25} \\ l_{31} l_{32} \ldots l_{35} \\ \vdots \quad \vdots \qquad \vdots \\ l_{51} l_{52} \ldots l_{55} \end{bmatrix}
$$

The intuition – part 3

Each factor f can explain the variation in each A by the estimated loading l_q. As we have more than one A there are more than one estimated f and l.

$$\mathbf{f} = \begin{bmatrix} \mathrm{f}_1 \\ \mathrm{f}_2 \\ \mathrm{f}_3 \\ \vdots \\ \mathrm{f}_5 \end{bmatrix} \; and \; \boldsymbol{l} = \begin{bmatrix} l_{11} l_{12} \dots l_{15} \\ l_{21} l_{22} \dots l_{25} \\ l_{31} l_{32} \dots l_{35} \\ \vdots \quad \vdots \qquad \vdots \\ l_{51} l_{52} \dots l_{55} \end{bmatrix}$$

So, how do we reduce the dimensionality of this information and identify the relevant factors?

Eigenvalues (λ)

A number such that a given matrix minus that number times the identity matrix has a zero determinant. <u>It conceptually represents that amount of variance accounted for by a factor</u>.

OK, but what does it mean?

Principal Component Analysis -> involves extracting linear composites of observed variables.

Factor Analysis -> formal model predicting observed variables from theoretical latent factors.

Usually produce similar results.

Generally speaking:

1. Run Principal Component Analysis if you want to simply reduce your correlated observed variables to a smaller set of important independent composite variables.

2. Run Factor Analysis if you assume or wish to test a theoretical model of latent factors causing observed variables.

OK, but what does it mean?

Reducing data dimensionality or measuring latent constructs.

Principal Component Analysis: each principal component is the linear combination of x-variables that has maximum variance (among all linear combinations). It accounts for as much remaining variation in the data as possible.

Factor Analysis is a method for modeling observed variables, and their covariance structure, in terms of a smaller number of underlying unobservable (latent) "factors."

In **PCA,** all of the observed variance is analyzed**,** while in **Factor Analysis** it is only the shared variances that is analyzed.

Source: PennState STAT 505 Online Course

Some details not covered here

Both PCA and Factor Analysis have important details that should be considered.

- Model assumptions;
- Goodness-of-fit;
- All rotations types;
- Non-continuous variable (polychoric or tetrachoric).

Agenda

1) What is Factor Analysis and PCA?
2) How to run it on Stata or R? (More on this next week)
3) Interpreting the outputs.

How to run it on Stata or R?

Stata:

- factor *varlist, method(“pf, pcf, ipf, ml”) factors(#) - optional*
- *rotate, (“varimax or promax”)   #blanks()#*
- scree
- predict f1 f2 f3 … fn #whatever name you want to each factor

R

- library(psych)
- fa(r = DATA, nfactors = #Factors, rotate = “Rotation", covar = “Correlation or Covariance matrix”, fm = “Estimation procedure", scores = "regression")
- ev <- eigen(“correlation matrix”)
- plot(ev$values, las = 1, type = "b") – scree plot

Agenda

1) What is Factor Analysis and PCA?

2) How to run it on Stata or R? (More on this next week)

3) Interpreting the outputs.

Interpreting the outputs

```
use https://s        at/stata/output/m255, clear

factor item13-item24, ipf factor(3)
(obs=1365)
```

iterated principal factor

Number of factors to be retained

Variables used

principal factors; 3 factors retained)

Factor	Eigenvalue[a]	Difference[b]	Proportion[c]	Cumulative[d]
1	5.85150	5.04464	0.8336	0.8336
2	0.80687	0.44540	0.1149	0.9485
3	0.36146	0.23001	0.0515	1.0000
4	0.13146	0.07619	0.0187	1.0187
5	0.05527	0.02362	0.0079	1.0266
6	0.03164	0.02946	0.0045	1.0311
7	0.00218	0.00658	0.0003	1.0314
8	-0.00440	0.01466	-0.0006	1.0308
9	-0.01906	0.02688	-0.0027	1.0281
10	-0.04594	0.01440	-0.0065	1.0215
11	-0.06035	0.03050	-0.0086	1.0129
12	-0.09084	.	-0.0129	1.0000

Source: https://stats.idre.ucla.edu/stata/output/factor-analysis/

Interpreting the outputs

```
use https://stats.idre.ucla.edu/st                          ar

factor item13-item24, ipf factor(3
(obs=1365)
```

Proportion of variance accounted for by this factor plus all of the previous ones

Variance explained by factor

(Iter ted principal factors; 3 fa ors retained)

Factor	Eigenvalue[a]	Difference[b]	Proportion[c]	Cumulative[d]
1	5.85150	5.04464	0.8336	0.8336
2	0.80687	0.44540	0.1149	0.9485
3	0.36146	0.23001	0.0515	1.0000
4	0.13146	0.07619	0.0187	1.0187
5	0.05527	0.02362	0.0079	1.0266
6	0.03164	0.02946	0.0045	1.0311
7	0.00218	0.00658	0.0003	1.0314
8	-0.00440	0.01466	-0.0006	1.0308
9	-0.01906	0.02688	-0.0027	1.0281
10	-0.04594	0.01440	-0.0065	1.0215
11	-0.06035	0.03050	-0.0086	1.0129
12	-0.09084	.	-0.0129	1.0000

Source: https://stats.idre.ucla.edu/stata/output/factor-analysis/

Represents both how the variables are weighted for each factor but also the correlation between the variables and the factor

	Factor Loadings[e]			
Variable	1	2	3	Uniqueness[f]
item13	0.71339	-0.39873	0.09231	0.32356
item14	0.70320	-0.33908	0.09782	0.38097
item15	0.72122	-0.24499	0.10575	0.40864
item16	0.64779	-0.18905	0.11144	0.53220
item17	0.78307	-0.07337	0.06670	0.37698
item18	0.73947	0.34478	0.11291	0.32157
item19	0.61655	0.41588	0.15515	0.42284
item20	0.55009	0.23916	0.09318	0.63152
item21	0.73173	0.11683	0.00067	0.45093
item22	0.61281	0.26089	-0.02282	0.55588
item23	0.81937	-0.02620	-0.34543	0.20863
item24	0.69515	0.01825	-0.38727	0.36646

Source: https://stats.idre.ucla.edu/stata/output/factor-analysis/

Error: proportion of the common variance of the variable not associated with the factors

A note on rotations:

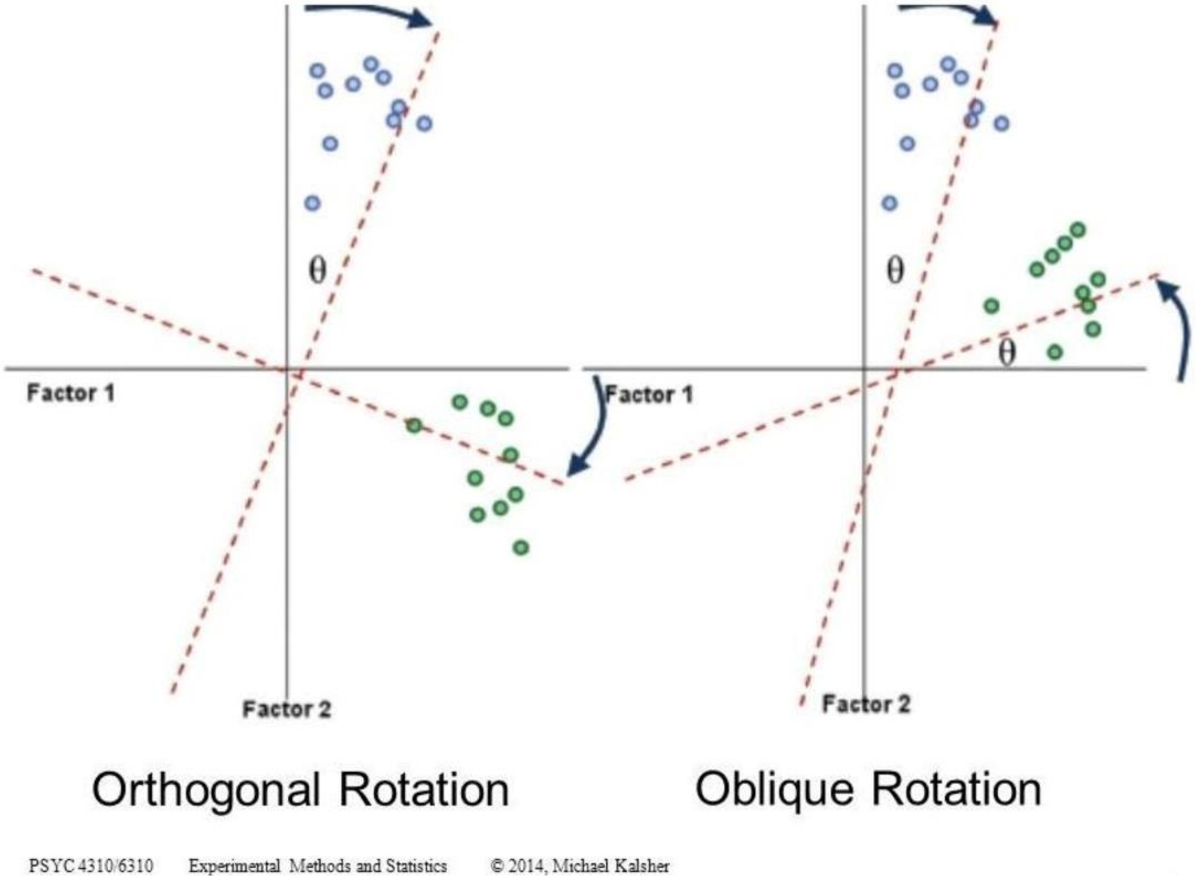
Rotations are done for the sake of interpretation of the extracted factors in factor analysis (or components in PCA).

It does not change the position of variables relative to each other in the space of the factors, i.e. correlations between variables are being preserved.

What are changed are the coordinates of the variable vectors' end-points onto the factor axes - the loadings.

- Orthogonal rotation (varimax): The factors are uncorrelated

- Oblique rotation (promax): The factors may be correlated



Source: https://stats.stackexchange.com/questions/151653/what-is-the-intuitive-reason-behind-doing-rotations-in-factor-analysis-pca-how

Rotation, varimax method: An orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. This method simplifies the interpretation of the factors.

```
rotate, varimax

Factor analysis/cor
    Method: iterated principal factors              Retained factors =
    Rotation: orthogonal varimax (Horst on)         Number of params =        33
```

Factor	Variance	Difference	Proportion	Cumulative
Factor1	2.94943	0.29428	0.4202	0.4202
Factor2	2.65516	1.23992	0.3782	0.7984
Factor3	1.41524	.	0.2016	1.0000

LR test: independent vs. saturated: chi2(66) = 8683.10 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

Variable	Factor1	Factor2	Factor3	Uniqueness
item13	0.7714			0.3236
item14	0.7256			0.3810
item15	0.6756			0.4086
item16	0.5908			0.5322
item17	0.5867	0.4461		0.3770
item18		0.7386		0.3216
item19		0.7281		0.4228
item20		0.5396		0.6315
item21	0.4020	0.5333	0.3210	0.4509
item22		0.5584		0.5559
item23	0.4488	0.3769	0.6692	0.2086
item24	0.3235	0.3205	0.6528	0.3665

(blanks represent abs(loading)<.3)

Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.6584	0.6121	0.4381
Factor2	-0.6840	0.7294	0.0088
Factor3	0.3141	0.3055	-0.8989

Source: https://stats.idre.ucla.edu/stata/output/factor-analysis/

Promax Rotation. An oblique rotation, which allows factors to be correlated. This rotation can be calculated more quickly than other rotations, so it is useful for large datasets.

```
rotate, promax

Factor analysis/cor
    Method: iterated principal factors              Retained factors =      3
    Rotation: oblique promax (Horst on)              Number of params =     33
```

Factor	Variance	Proportion	Rotated factors are correlated
Factor1	4.86265	0.6927	
Factor2	4.52052	0.6440	
Factor3	4.30842	0.6138	

LR test: independent vs. saturated: chi2(66) = 8683.10 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances[i]

Variable	Factor1	Factor2	Factor3	Uniqueness[j]
item13	0.8518			0.3236
item14	0.7855			0.3810
item15	0.6969			0.4086
item16	0.6044			0.5322
item17	0.5087			0.3770
item18		0.7626		0.3216
item19		0.8200		0.4228
item20		0.5541		0.6315
item21		0.4298		0.4509
item22		0.5265		0.5559
item23			0.7187	0.2086
item24			0.7502	0.3665

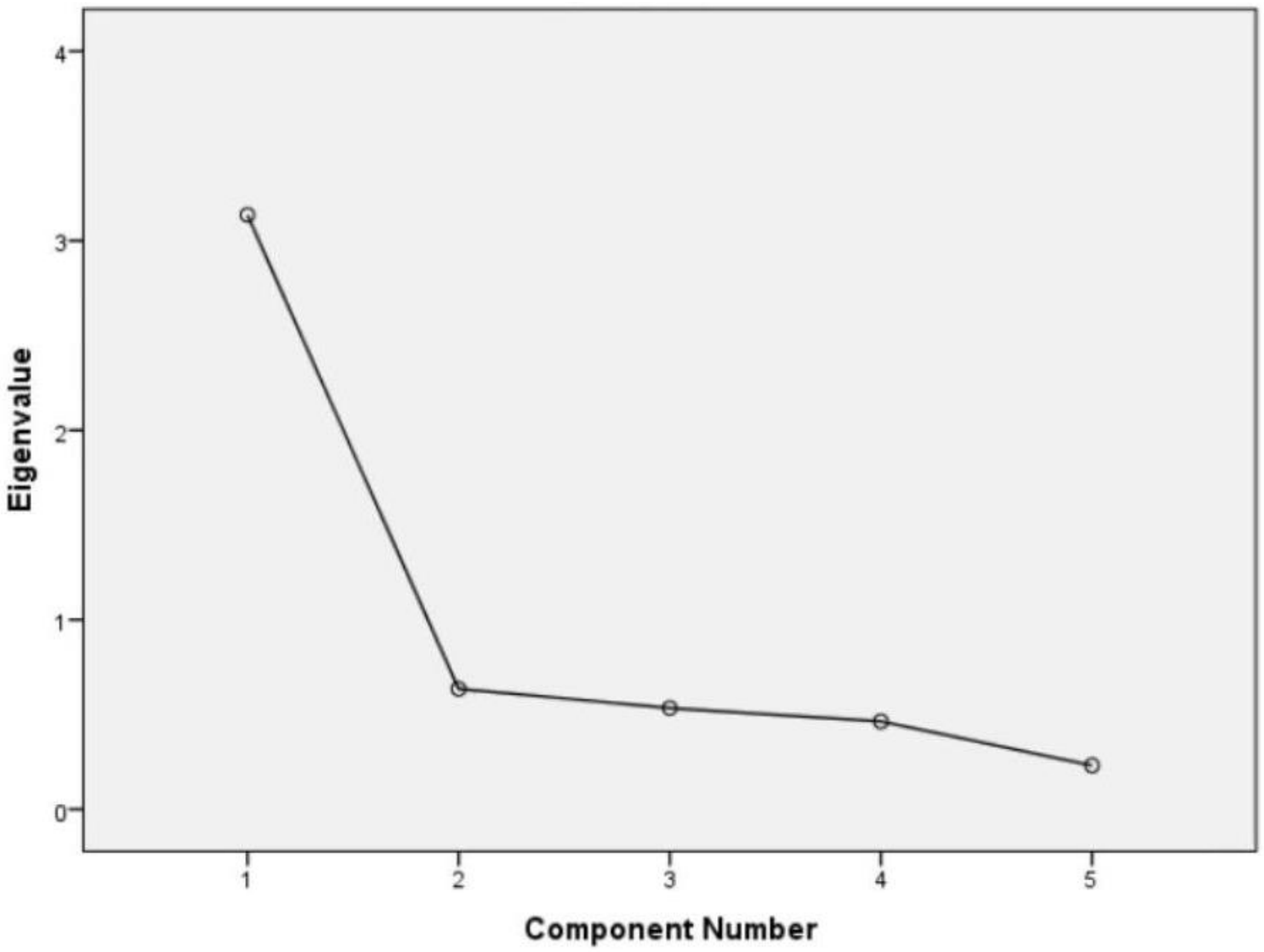
(blanks represent abs(loading)<.3)

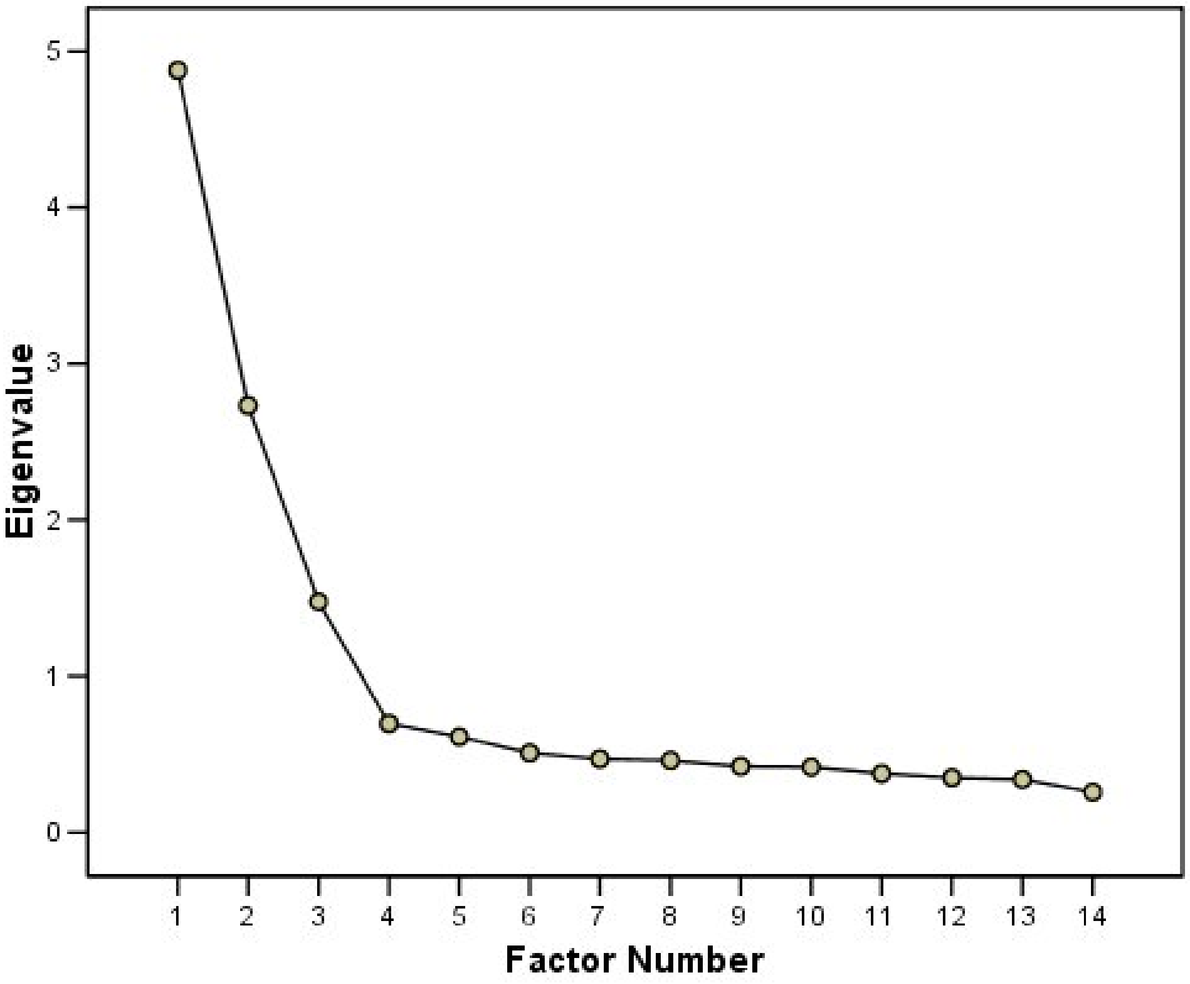
Factor rotation matrix

	Factor1	Factor2	Factor3
Factor1	0.8977	0.8593	0.8479
Factor2	-0.4157	0.4864	0.0071
Factor3	0.1462	0.1581	-0.5301

Source: https://stats.idre.ucla.edu/stata/output/factor-analysis/

What is a scree plot? (More next week)





Next week, practice round!

Thank you!

filiperecch@Stanford.edu