R WORKSHOP – DAY 2

Randi L. Garcia

Smith College

July 17th, 19th, and 21st

DAY 2

- ANOVA and regression
- Preparing APA style manuscripts
- Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA)
- Path Analysis and SEM

ANOVA and Regression

- **Analysis of Variance (ANOVA)** is used to compare the means of a numerical variable across levels of a categorical variable (3+ levels)
  - Only 2 levels, what test do we use?

- **Simple Linear Regression (SLR)** is used to find the relationship between one numerical predictor variable and one numerical response (outcome or DV) variable.

- **Multiple Regression** is used to find the relationship between predictor and response controlling for other variables.

ANOVA and Regression

- **Logistic Regression** is used to model the probability of being in a certain group based on numerical predictors.
  - i.e., The response variable is dichotomous
  - This is called a **Generalized Linear Model (GLM)**

- **χ^2-Test (Chi-squared Test)** is used to test if two categorical variables are associated.
  - For example, is the distribution of education levels more skewed towards higher degrees for men than for women?

ANOVA and Regression

Explanatory **(IV or predictor)**	**Response** **(DV or outcome variable)**	
	Numerical	Categorical (2 levels: dichotomous)
Categorical (levels = 2)	t-Test	χ^2-Test (two-prop test)
1 Numerical	SLR	Logistic Regression
Categorical (levels >= 3)	ANOVA	χ^2-Test
2 or more Numerical	Multiple Regression	Logistic Regression

ANOVA and Regression

Inference Test	R function
t-Test	t.test()
ANOVA	aov()
SLR and Multiple Regression	lm()
χ^2-Test	chisq.test()
Logistic Regression	glm()

R MARKDOWN FILE

PREPARING APA STYLE MANUSCRIPTS

Connie Zhang, 19’ and Emma Ning, 19’

Exploratory Factor Analysis (EFA)

- Often we want to be able to describe a relatively large number of **items** by a much fewer number of **factors**.

- In the bfi dataset there are 25 items measuring personality, but are there just a few underlying factors that are responsible for people's scores on those items?

- We might guess what those are (e.g., extroversion, conscientiousness, etc.), but if we didn't know we could use **EFA** to let the data tell us about the underlying dimensions.

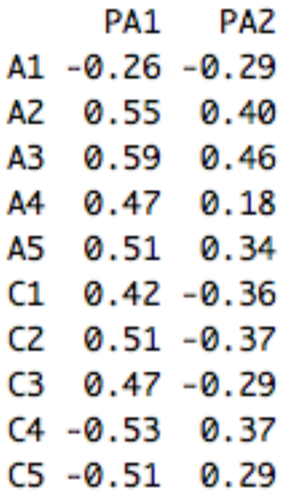
Exploratory Factor Analysis (EFA)

- Exploratory Factor Analysis (EFA) will use inter-correlations among the items to give us a sense of…
  1. how many factors may be present,
  2. which items can be explained by which factors, and
  3. the extent to which these underlying factors are correlated with each other.

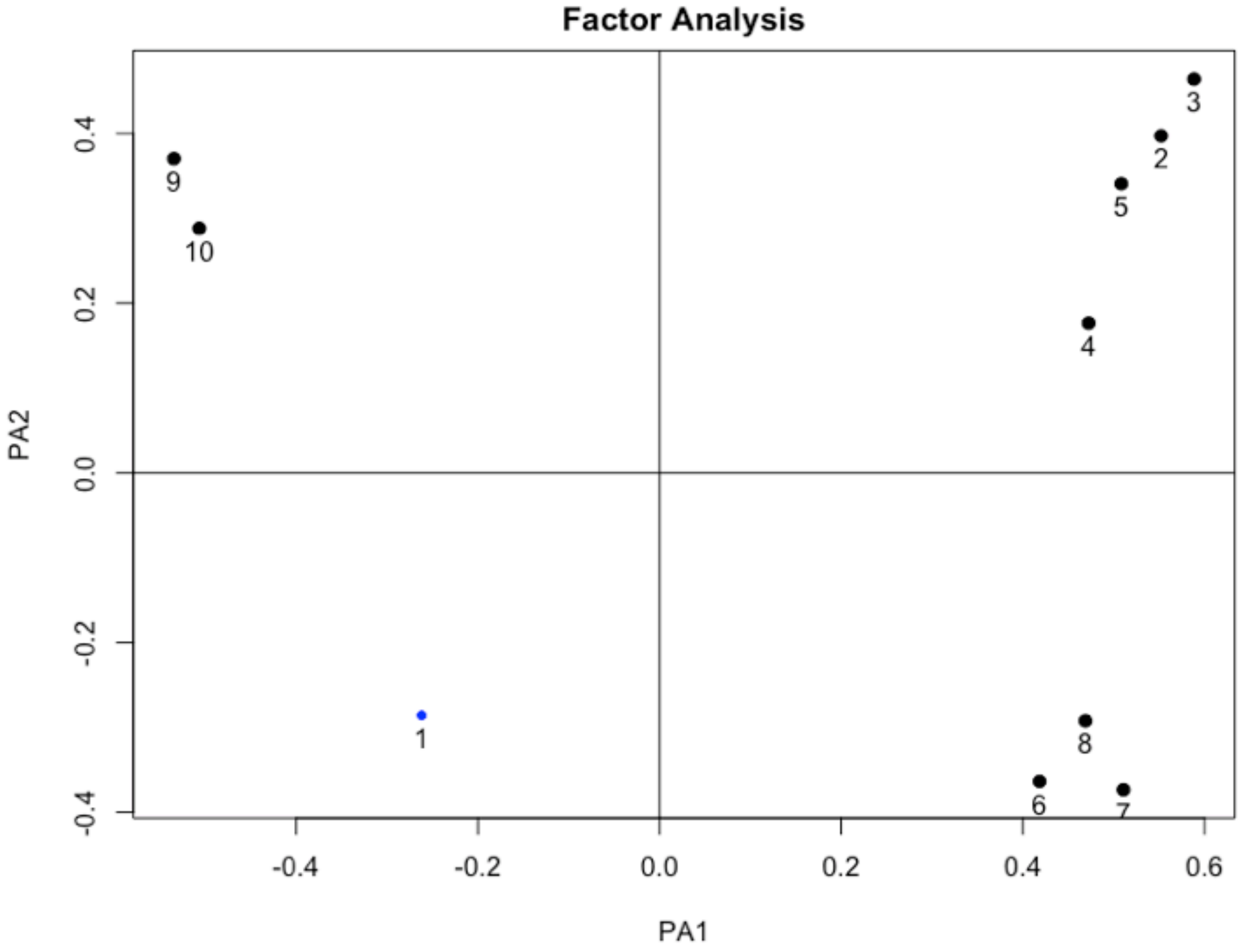
- EFA is just that, exploratory.
  - It is important to keep in mind that in the end this is a data driven technique. Meaning that peculiarities in the data may lead you to a rather weird solution.
  - It takes some sense finesse, listen to what your data is telling you.

Factor Rotation

- Unrotated solution

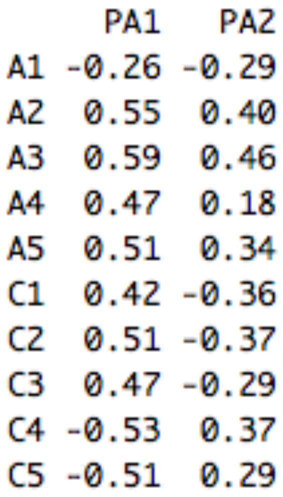


```
      PA1   PA2
A1 -0.26 -0.29
A2  0.55  0.40
A3  0.59  0.46
A4  0.47  0.18
A5  0.51  0.34
C1  0.42 -0.36
C2  0.51 -0.37
C3  0.47 -0.29
C4 -0.53  0.37
C5 -0.51  0.29
```

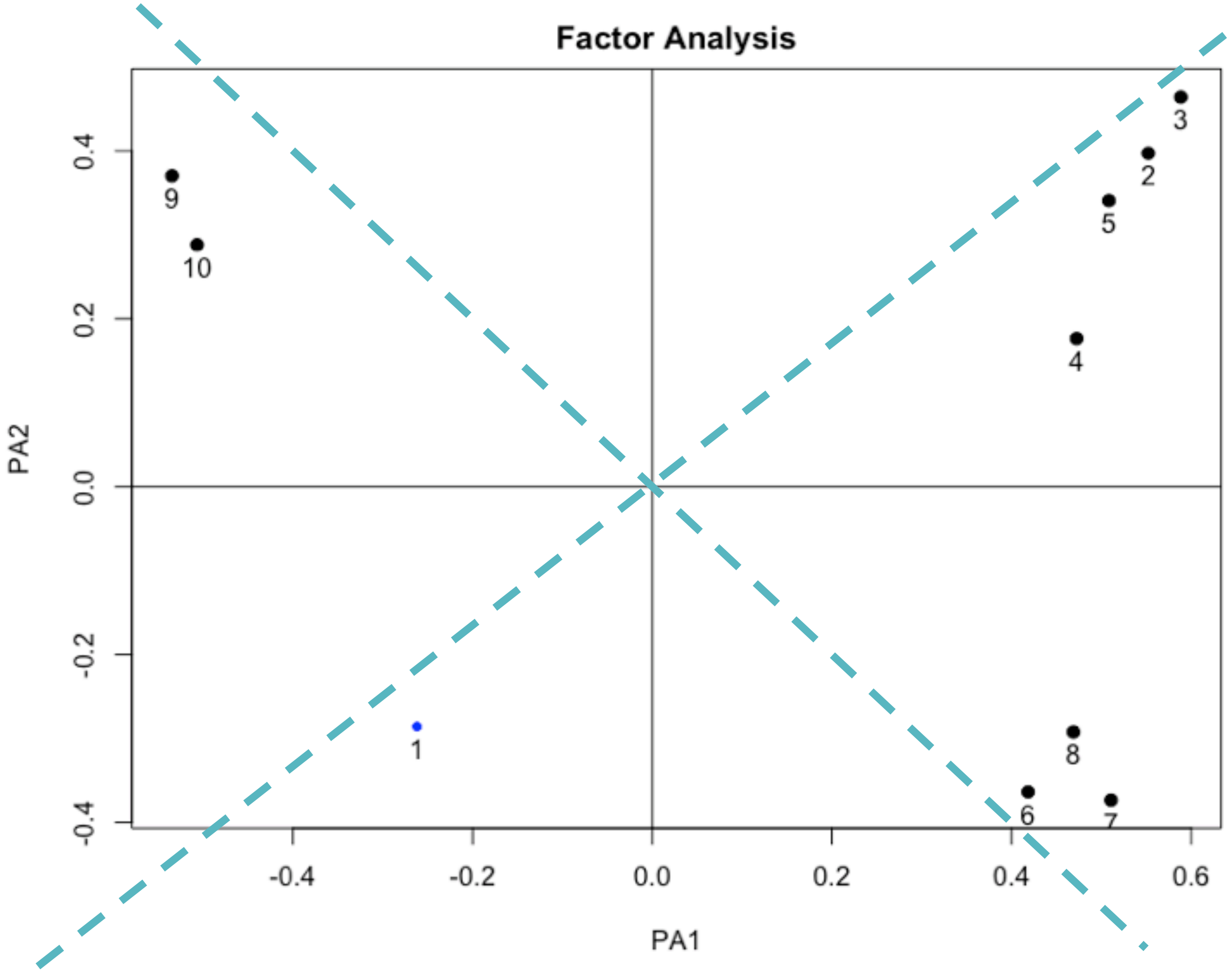


Factor Rotation

- Unrotated solution

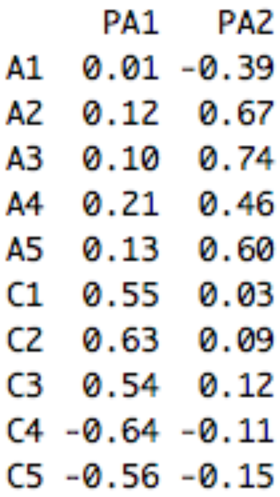


```
      PA1   PA2
A1  -0.26 -0.29
A2   0.55  0.40
A3   0.59  0.46
A4   0.47  0.18
A5   0.51  0.34
C1   0.42 -0.36
C2   0.51 -0.37
C3   0.47 -0.29
C4  -0.53  0.37
C5  -0.51  0.29
```

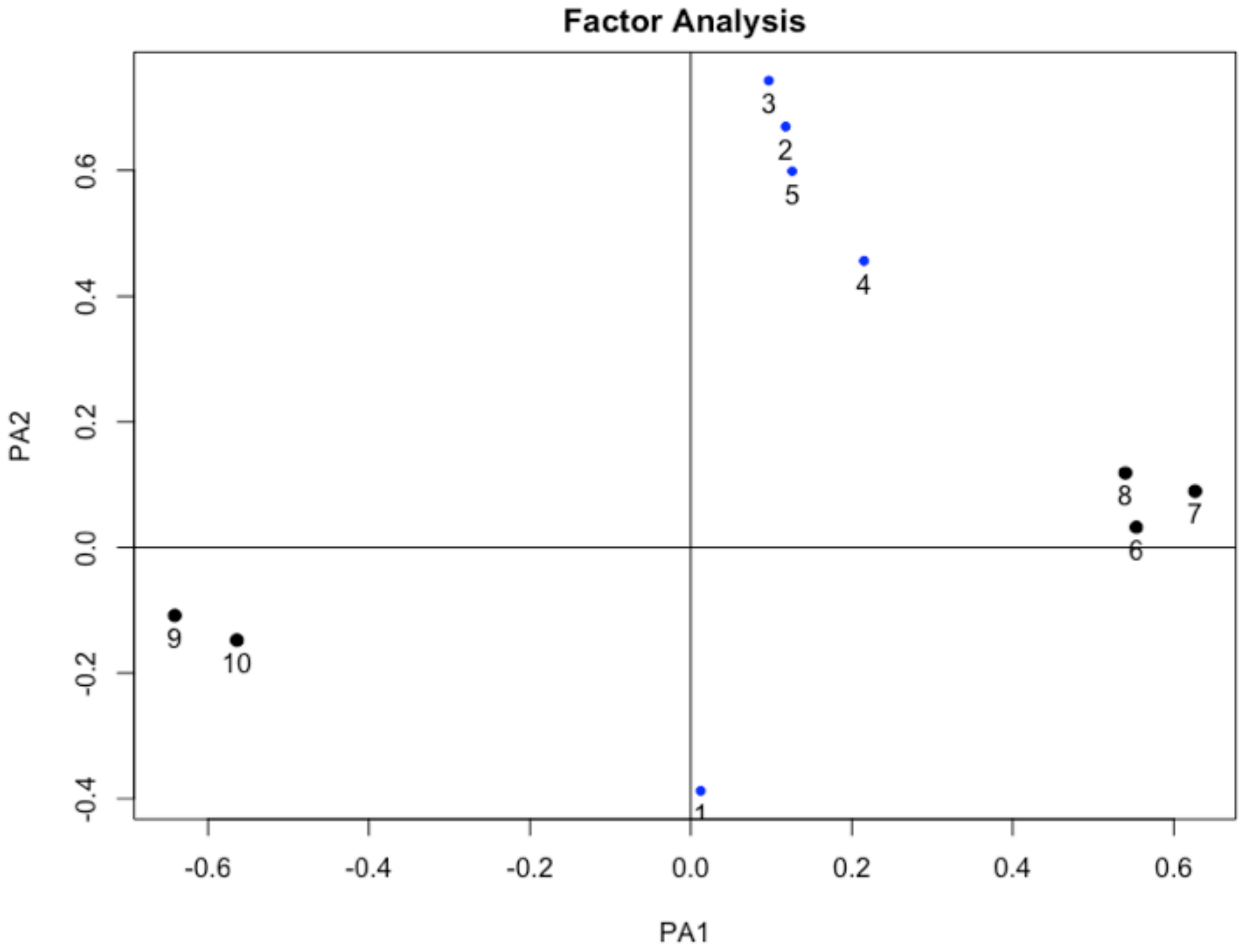


Factor Rotation

- Orthogonal rotation

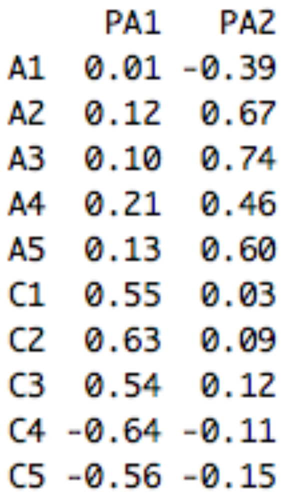


```
     PA1   PA2
A1  0.01 -0.39
A2  0.12  0.67
A3  0.10  0.74
A4  0.21  0.46
A5  0.13  0.60
C1  0.55  0.03
C2  0.63  0.09
C3  0.54  0.12
C4 -0.64 -0.11
C5 -0.56 -0.15
```

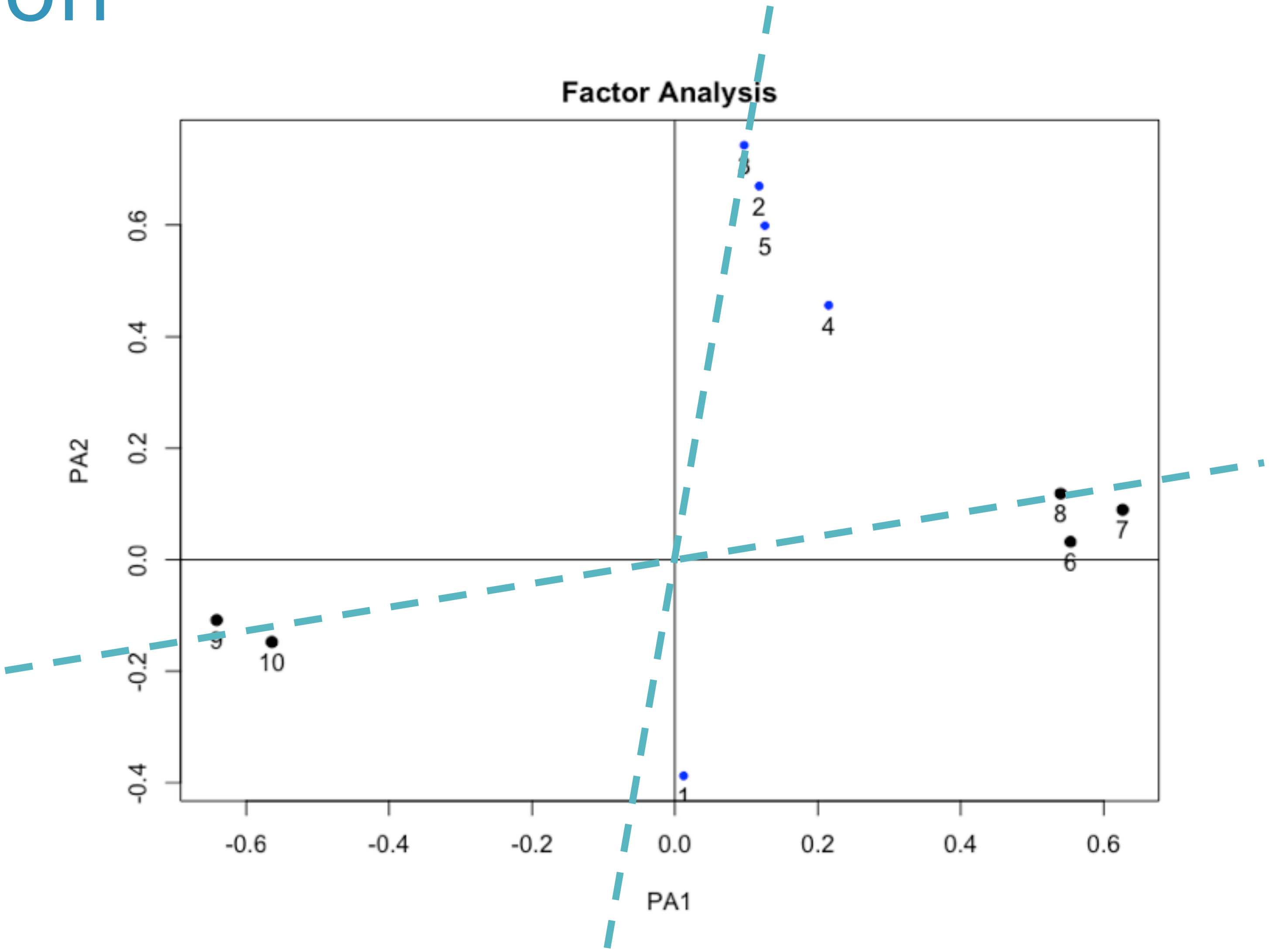


Factor Rotation

- Orthogonal rotation

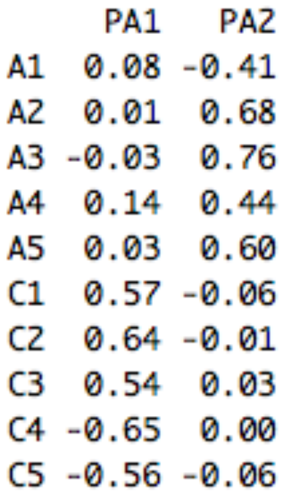


```
     PA1   PA2
A1  0.01 -0.39
A2  0.12  0.67
A3  0.10  0.74
A4  0.21  0.46
A5  0.13  0.60
C1  0.55  0.03
C2  0.63  0.09
C3  0.54  0.12
C4 -0.64 -0.11
C5 -0.56 -0.15
```

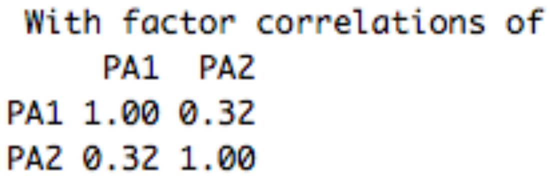


Exploratory Factor Analysis (EFA)

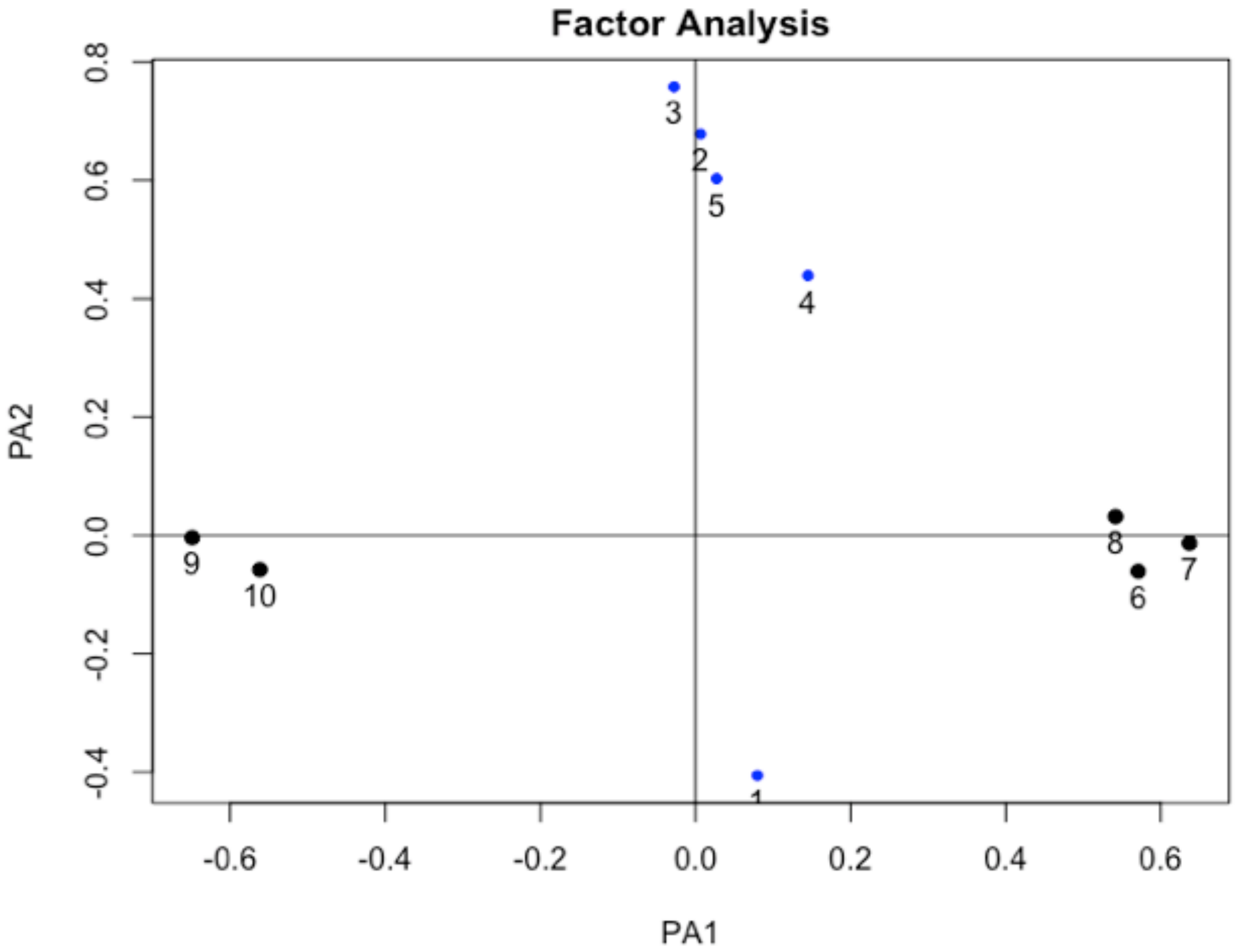
- Oblique factor rotation



```
      PA1   PA2
A1   0.08 -0.41
A2   0.01  0.68
A3  -0.03  0.76
A4   0.14  0.44
A5   0.03  0.60
C1   0.57 -0.06
C2   0.64 -0.01
C3   0.54  0.03
C4  -0.65  0.00
C5  -0.56 -0.06
```



```
 With factor correlations of
     PA1  PA2
PA1 1.00 0.32
PA2 0.32 1.00
```



Exploratory Factor Analysis (EFA)

- We will use the psych package

Inference Test	R function
Factor Analysis	fa()
Principal Component Analysis	principal()

R MARKDOWN FILE

Confirmatory Factor Analysis (CFA)

data(HolzingerSwineford1939)

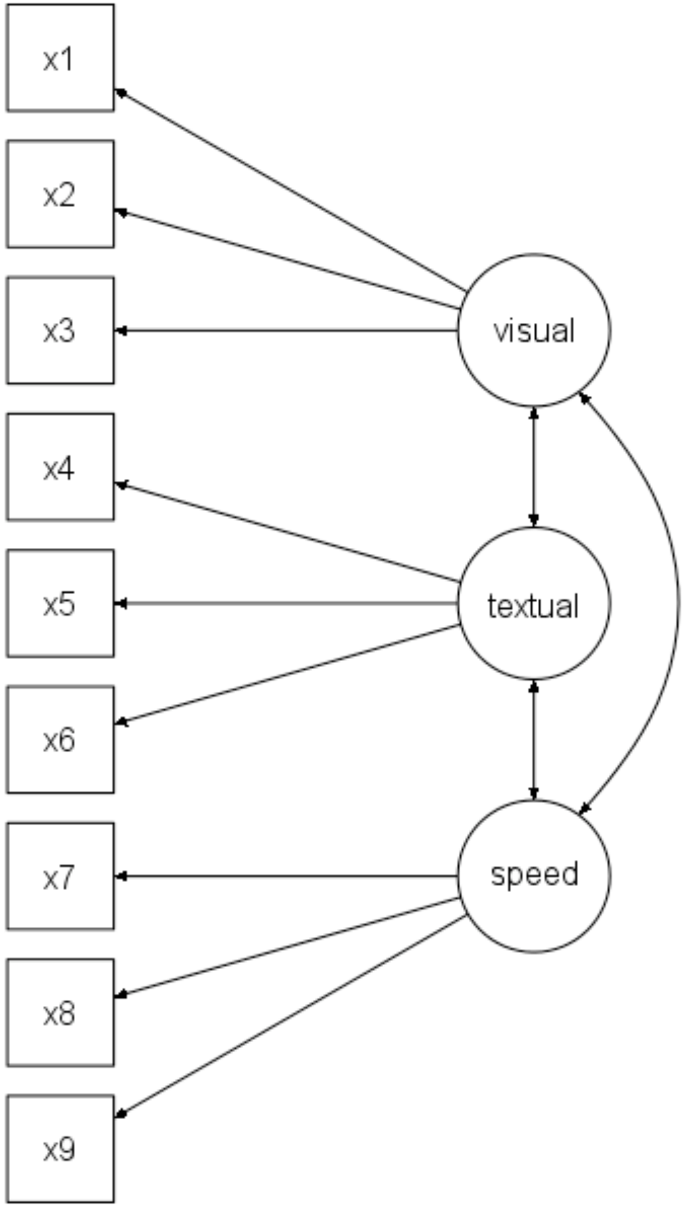
- Mental ability test score from 7th and 8th grade children from two schools
  - A *visual* factor measured by 3 variables: x1, x2 and x3
  - A *textual* factor measured by 3 variables: x4, x5 and x6
  - A *speed* factor measured by 3 variables: x7, x8 and x9

- We want to test if indeed these measures fall on these three scales as we hypothesize.
- We are *confirming* a hypothesized factor structure instead of exploring.

Visual factor: x1, x2 and x3

Textual factor: x4, x5 and x6

Speed factor: x7, x8 and x9

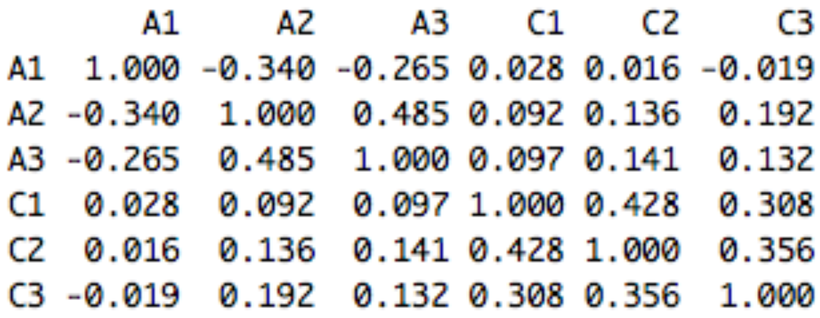


Confirmatory Factor Analysis (CFA)

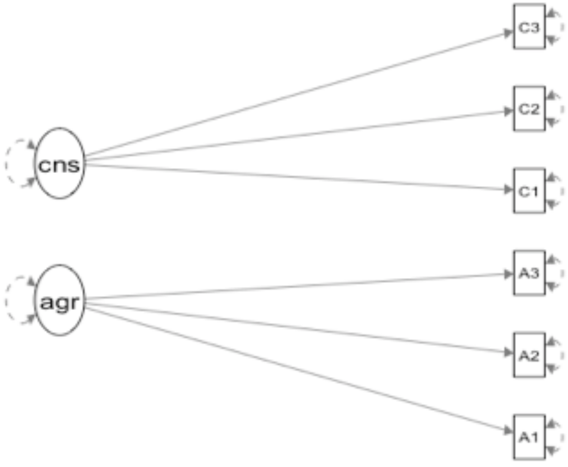
- Does the model we have in our heads actually fit the data?
    - Assessed with fit statistics

Model

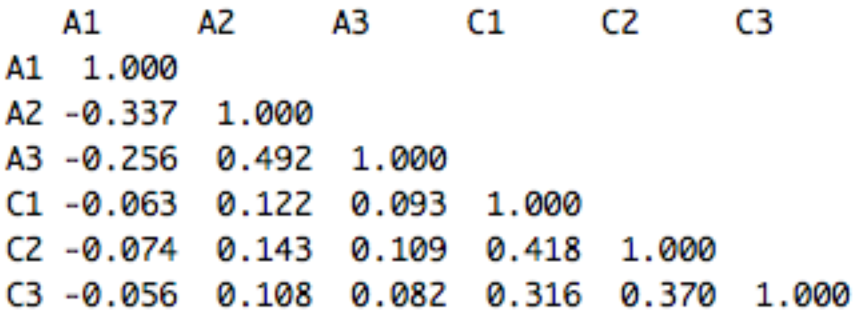
Data cor matrix



```
       A1     A2     A3    C1    C2     C3
A1  1.000 -0.340 -0.265 0.028 0.016 -0.019
A2 -0.340  1.000  0.485 0.092 0.136  0.192
A3 -0.265  0.485  1.000 0.097 0.141  0.132
C1  0.028  0.092  0.097 1.000 0.428  0.308
C2  0.016  0.136  0.141 0.428 1.000  0.356
C3 -0.019  0.192  0.132 0.308 0.356  1.000
```



Model implied cor matrix



```
      A1     A2     A3     C1     C2     C3
A1  1.000
A2 -0.337  1.000
A3 -0.256  0.492  1.000
C1 -0.063  0.122  0.093  1.000
C2 -0.074  0.143  0.109  0.418  1.000
C3 -0.056  0.108  0.082  0.316  0.370  1.000
```

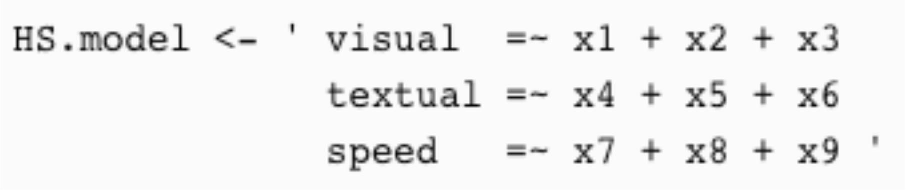
Fit?

Confirmatory Factor Analysis (CFA)

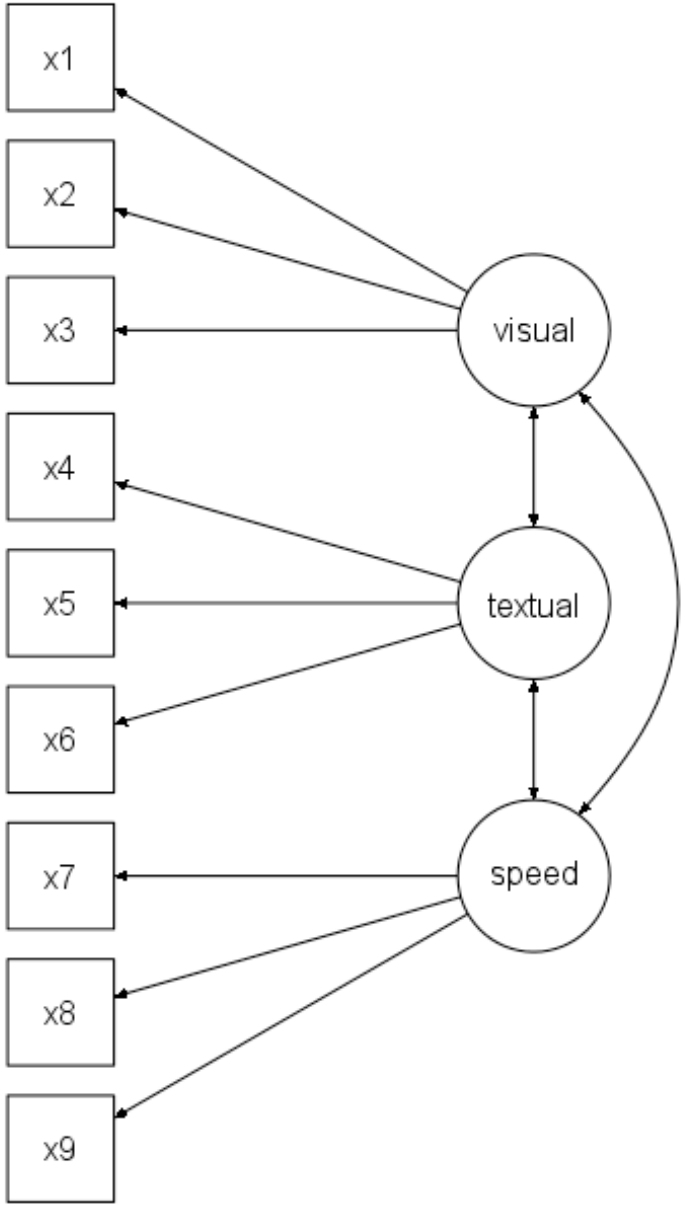
- We will use the R package lavaan to fit CFAs
  - most widely used **Structural Equation Modeling (SEM)** package in R.

- **Step 1:** Specify the model
- **Step 2:** Fit the model
- **Step 3:** Ask for the output you want

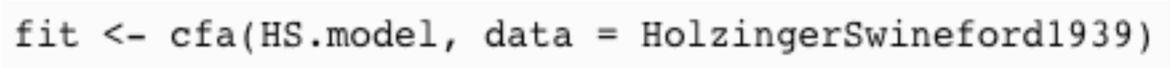
Step 1:
Specify the Model



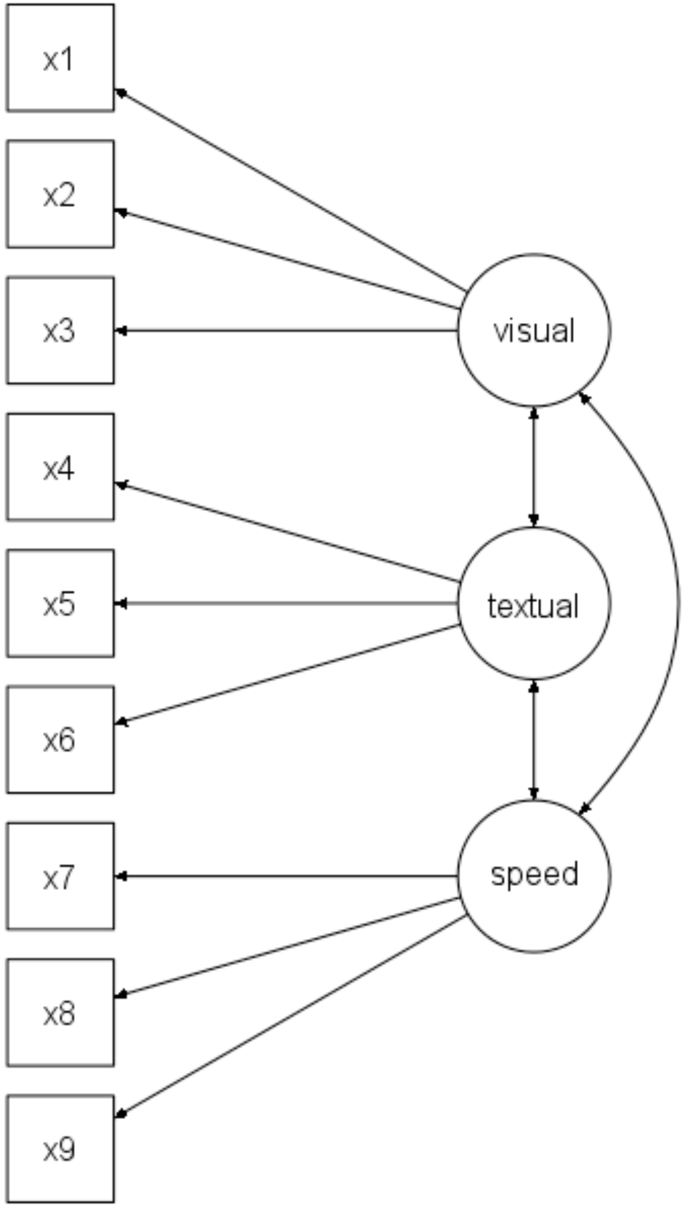
```
HS.model <- ' visual  =~ x1 + x2 + x3
              textual =~ x4 + x5 + x6
              speed   =~ x7 + x8 + x9 '
```



Step 2:
Fit the Model



```
fit <- cfa(HS.model, data = HolzingerSwineford1939)
```



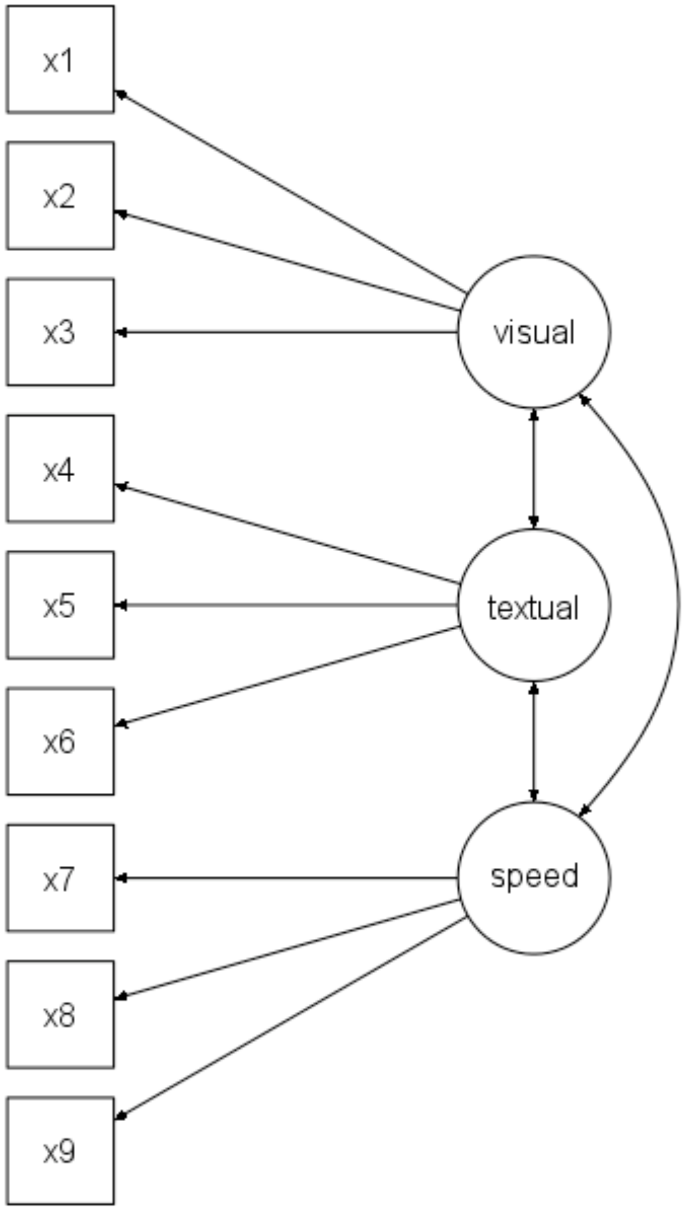
Step 3: Ask for the output you want

```
summary(fit, fit.measures = TRUE)
```

```
parameterEstimates(fit)
```

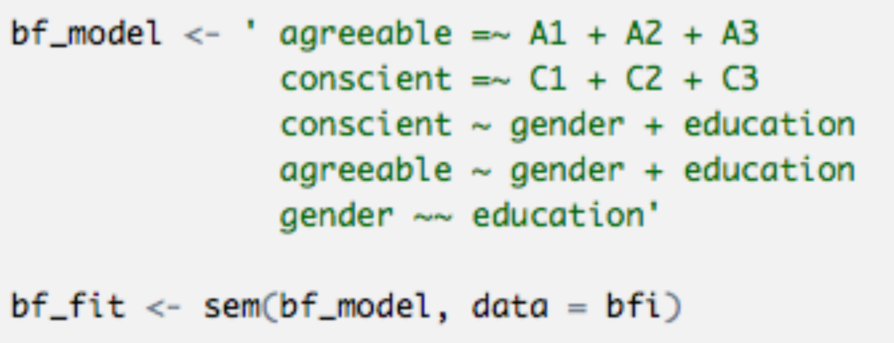
```
inspect(fit)
```

```
modindices(fit)
```



Path Analysis and SEM

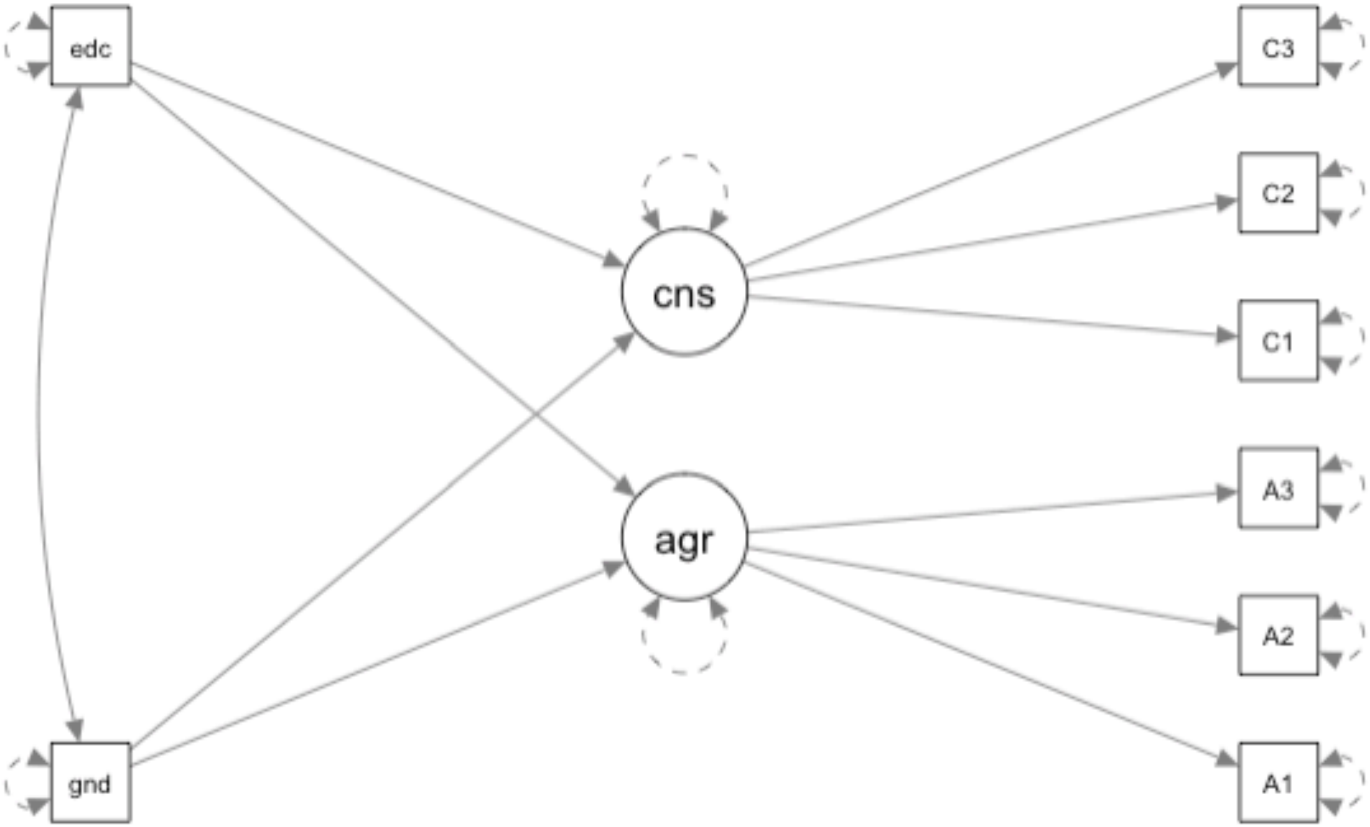
- Now we can add regression equations in the mix with our latent variables.
- We can use our latent variables as predictors (IVs) or as response variables (DVs).
- Simultaneously estimate multiple regression equations
  - A **multivariate data analysis** approach because we can have multiple response variables.
  - Think solving a system of equations!



```
bf_model <- ' agreeable =~ A1 + A2 + A3
              conscient =~ C1 + C2 + C3
              conscient ~ gender + education
              agreeable ~ gender + education
              gender ~~ education'

bf_fit <- sem(bf_model, data = bfi)
```

Path Analysis and SEM



R MARKDOWN FILE