Universiteit Utrecht

Theory Construction and Statistical Modeling

Welcome!

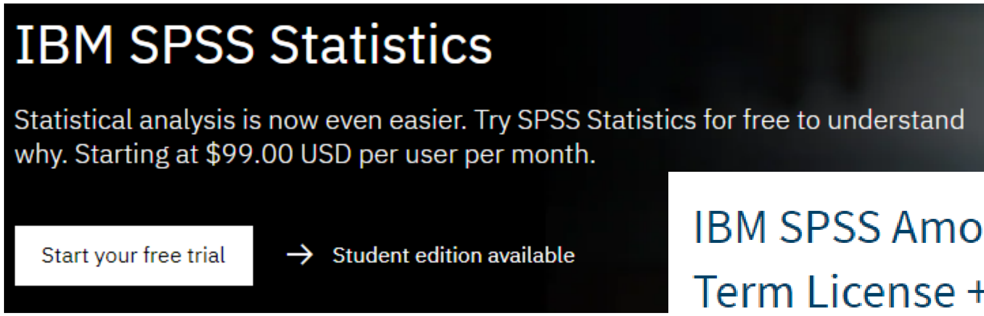
Lecturer

- Caspar van Lissa
  - Assistant professor
  - Lectures
  - Course coördinator

Benefits of using R

- It's free
- You can install it anywhere, even run online
- EVERY analysis is available in R
- Reproducible research
- Beautiful graphics
- Lots of help/support in online forums
- Easily interface with other programs

Before, we used SPSS and AMOS

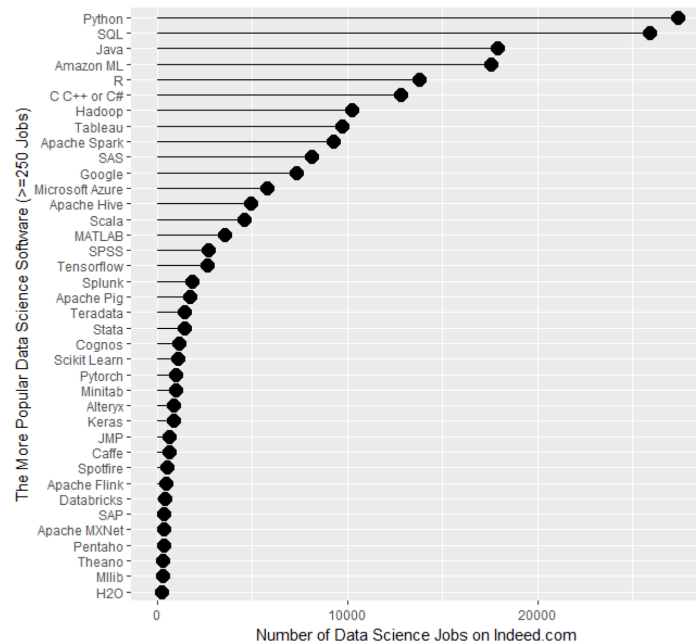


IBM SPSS Amos Concurrent User Initial Fixed Term License + SW Subscription & Support 12 Months

£1,719.00 excl VAT

- Expensive (even if you get a subsidized discount)
- Closed source, so hard to verify if calculations are correct
- Slow to develop new functionality

Before, we used SPSS and AMOS



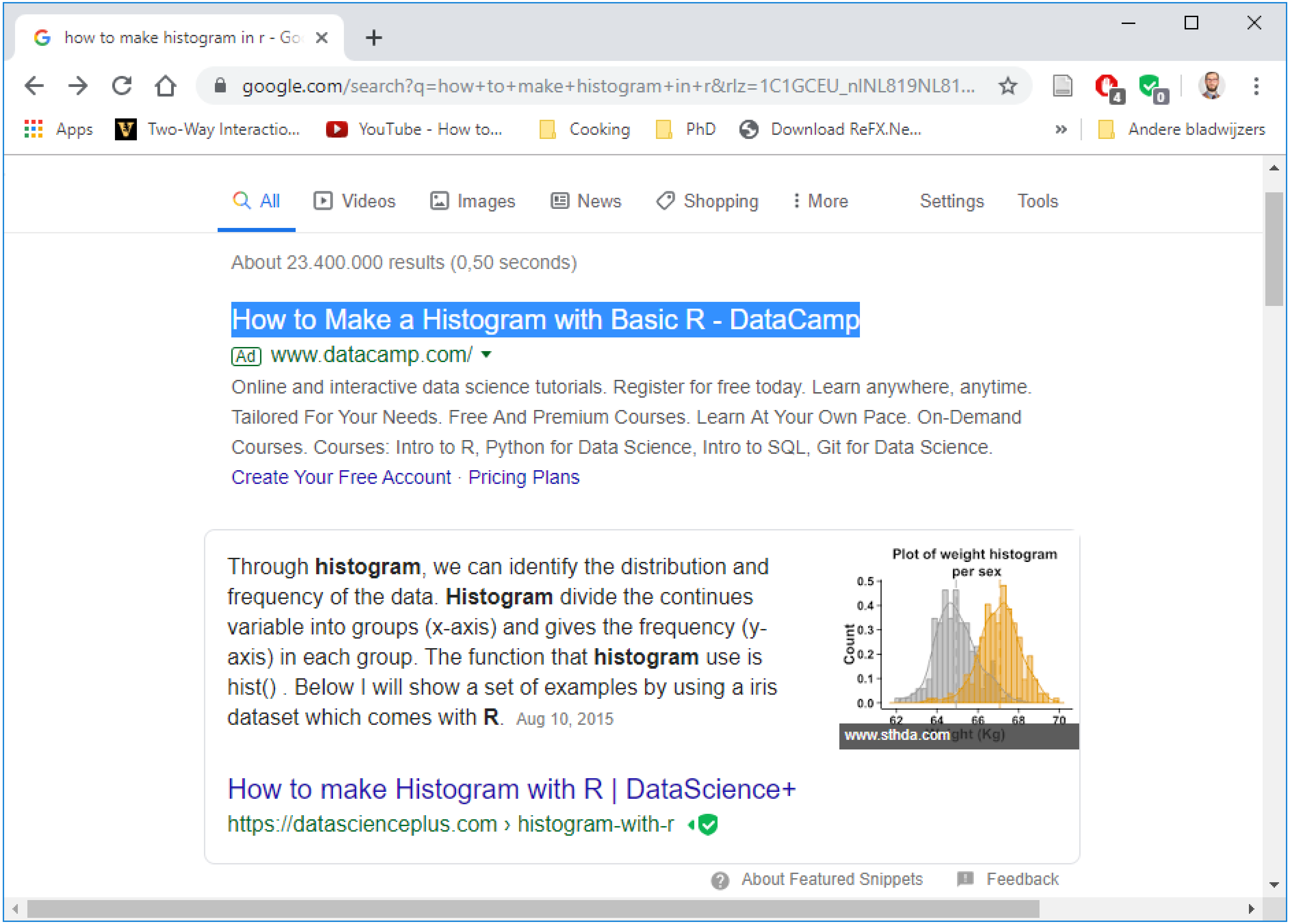
- Little demand on the job market (from r4stats.com)
- Large part of the course devoted to just dealing with the interface

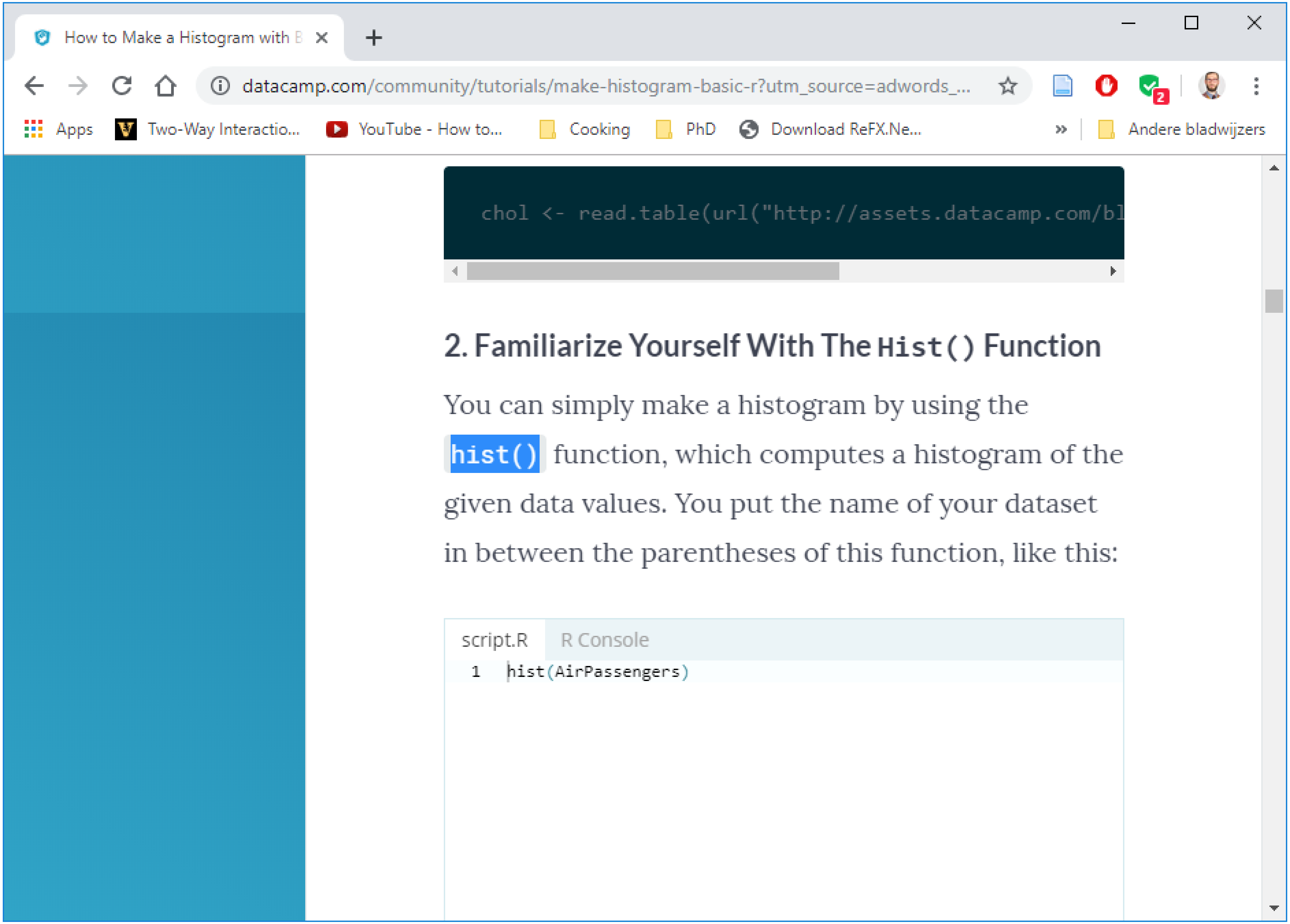
What does R give you?

- You have to learn new software anyway – why not something **useful**?
- Saves time
- Saves money
- It’s a real life/job skill
  - R can help you get jobs
  - R can make your job easier
- People will take you more seriously (seriously!)
- Develop logical reasoning skills
- High threshold: You can use R to automate many things (e.g., the GitBook is written in R)

Philosophy of “learning R”

- **Don’t** try to “learn R”
  - It’s too big
  - Everything you can think of is possible in R
  - Just focus on one task at a time
- Copy-paste code, then adapt it
  - From manual, from previous exercise, from internet
  - Make small changes when necessary
  - Check that it works as expected
- Learn and understand as you go
- Use the Help function in R: `?hist`, or select function and press F1
- Google “how to … in R” a LOT
  - E.g., “how to make histogram in R”
  - Stackexchange and R-bloggers are great


how to make histogram in r - Go
google.com/search?q=how+to+make+histogram+in+r&rlz=1C1GCEU_nlNL819NL81...
Apps
Two-Way Interactio...
YouTube - How to...
Cooking
PhD
Download ReFX.Ne...
Andere bladwijzers
All
Videos
Images
News
Shopping
More
Settings
Tools
About 23.400.000 results (0,50 seconds)
How to Make a Histogram with Basic R - DataCamp
Ad www.datacamp.com/
Online and interactive data science tutorials. Register for free today. Learn anywhere, anytime. Tailored For Your Needs. Free And Premium Courses. Learn At Your Own Pace. On-Demand Courses. Courses: Intro to R, Python for Data Science, Intro to SQL, Git for Data Science.
Create Your Free Account · Pricing Plans
Through histogram, we can identify the distribution and frequency of the data. Histogram divide the continues variable into groups (x-axis) and gives the frequency (y-axis) in each group. The function that histogram use is hist() . Below I will show a set of examples by using a iris dataset which comes with R. Aug 10, 2015
Plot of weight histogram per sex
Count
0.5
0.4
0.3
0.2
0.1
0.0
62
64
66
68
70
www.sthda.com
How to make Histogram with R | DataScience+
https://datascienceplus.com › histogram-with-r
About Featured Snippets
Feedback


How to Make a Histogram with B
datacamp.com/community/tutorials/make-histogram-basic-r?utm_source=adwords_...
Apps
Two-Way Interactio...
YouTube - How to...
Cooking
PhD
Download ReFX.Ne...
Andere bladwijzers
chol <- read.table(url("http://assets.datacamp.com/bl
2. Familiarize Yourself With The Hist() Function
You can simply make a histogram by using the hist() function, which computes a histogram of the given data values. You put the name of your dataset in between the parentheses of this function, like this:
script.R
R Console
1 hist(AirPassengers)

WHAT IS A MODEL?

What is a model?

- A schematic description of a system, theory, or phenomenon that accounts for its known or inferred properties and may be used for further study of its characteristics
- A simplified description, especially a mathematical one, of a system or process, to assist calculations and predictions
- Clearly defined level of analysis, operationalized variables, specified relationships between variables (see: Smaldino)

What will you learn in TCSM?

- How to translate a (verbal) social scientific **theory** into a statistical **model**
- How to analyze your data with these models
- How to interpret and report your results

What will you learn in TCSM?

- How to test a specific theory with an appropriate statistical model (**Confirmatory)**
- How to fit the model to data and reflect back to substantive theory? (**Exploratory)**

- Starting point:
  - Already have an appropriate (quantitative) dataset
  - Focus on **data analysis: modeling**

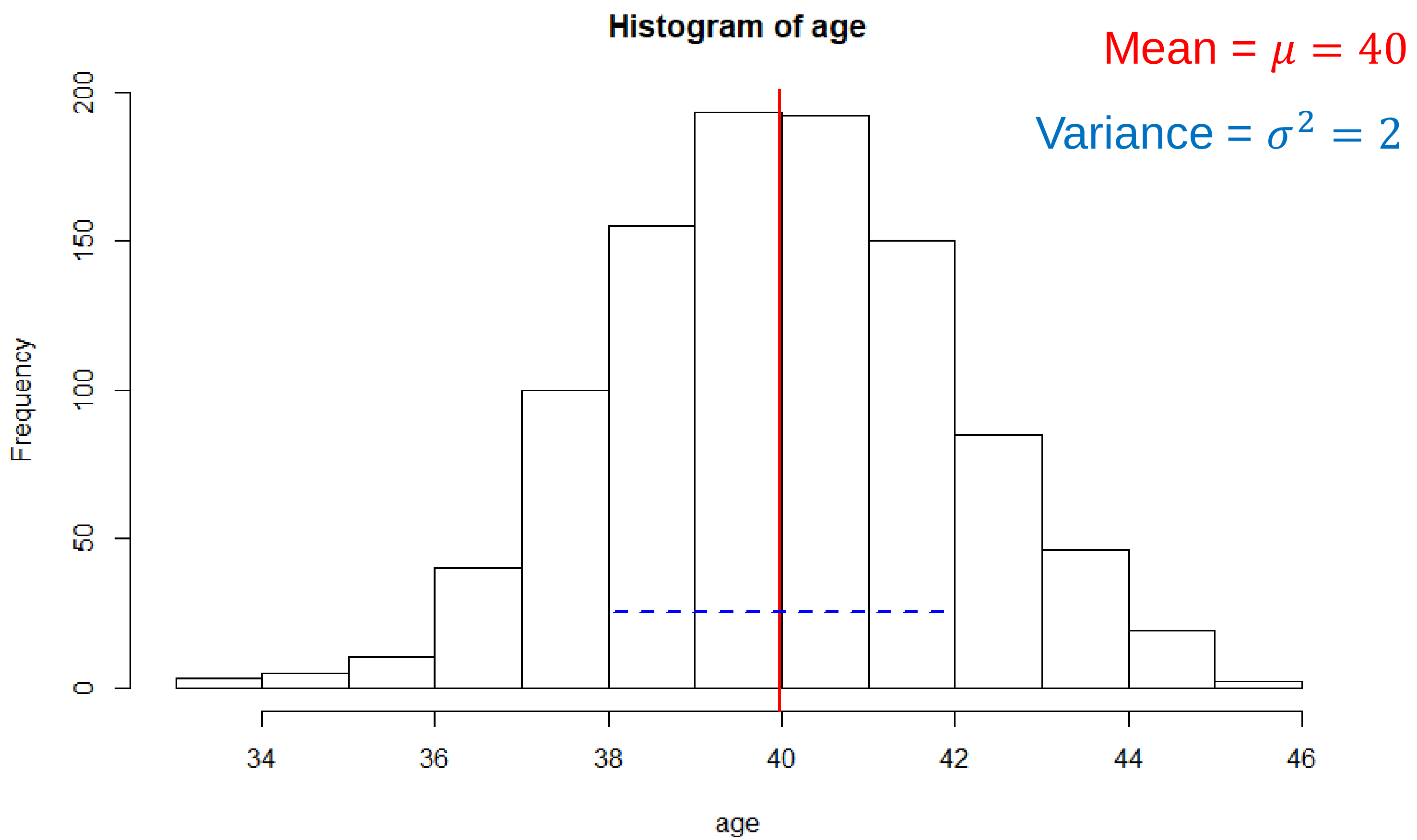
Statistical Model

- Explicates our ideas about how our observed data was generated
- We have observed **variables** with certain **characteristics**
- We develop a model to explain those characteristics

Variables and Characteristics

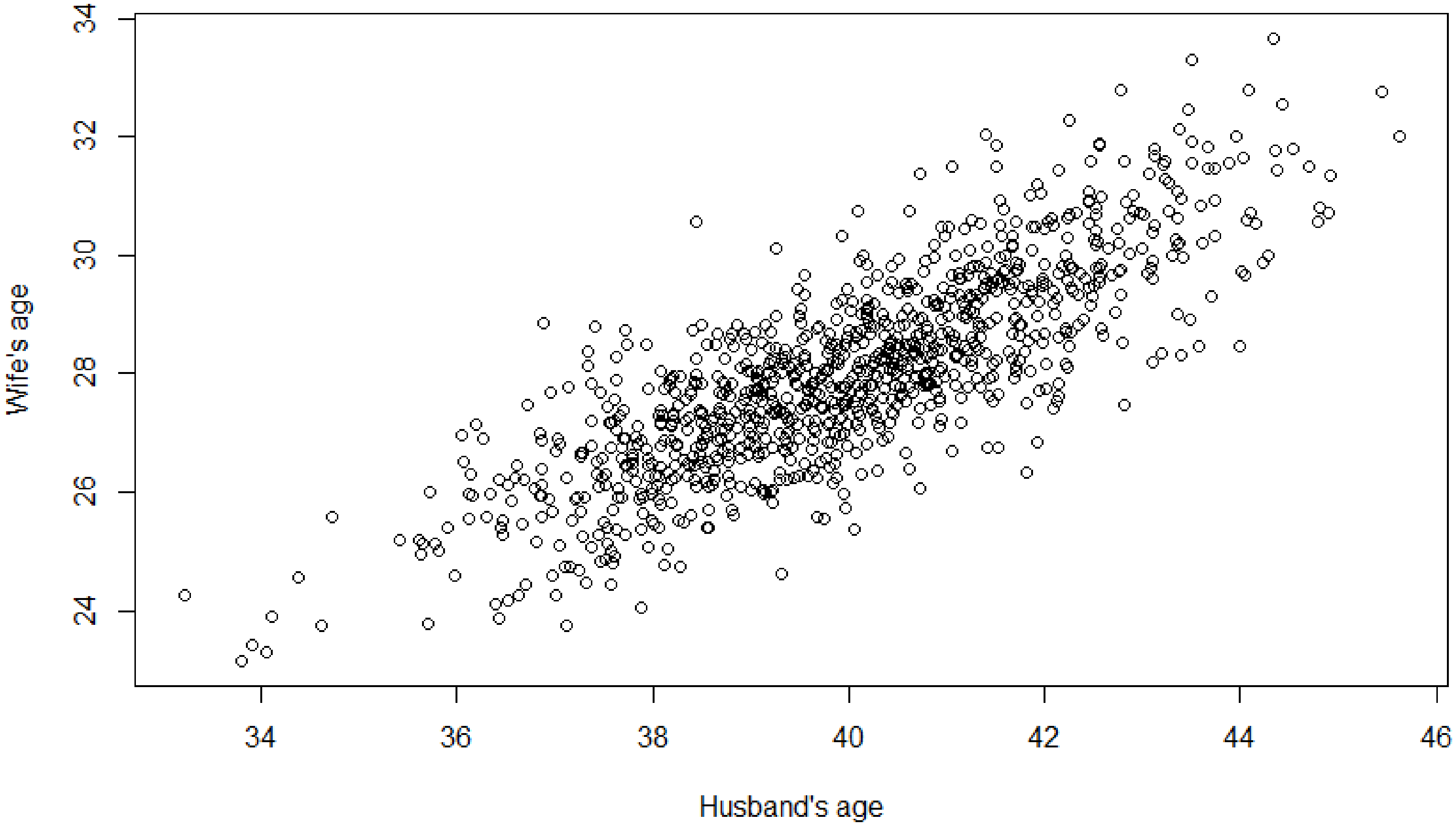
- What is a variable?
  - Anything that can **vary** i.e., take on different values
  - Height, age, intelligence, diagnosis, happiness…
- What are possible characteristics of variables?
  - By themselves (univariate): mean, variance
  - How they relate to another variable (bivariate): correlation/covariance

Univariate



Histogram of age
Mean = $\mu = 40$
Variance = $\sigma^2 = 2$
Frequency
0
50
100
150
200
34
36
38
40
42
44
46
age

Bivariate

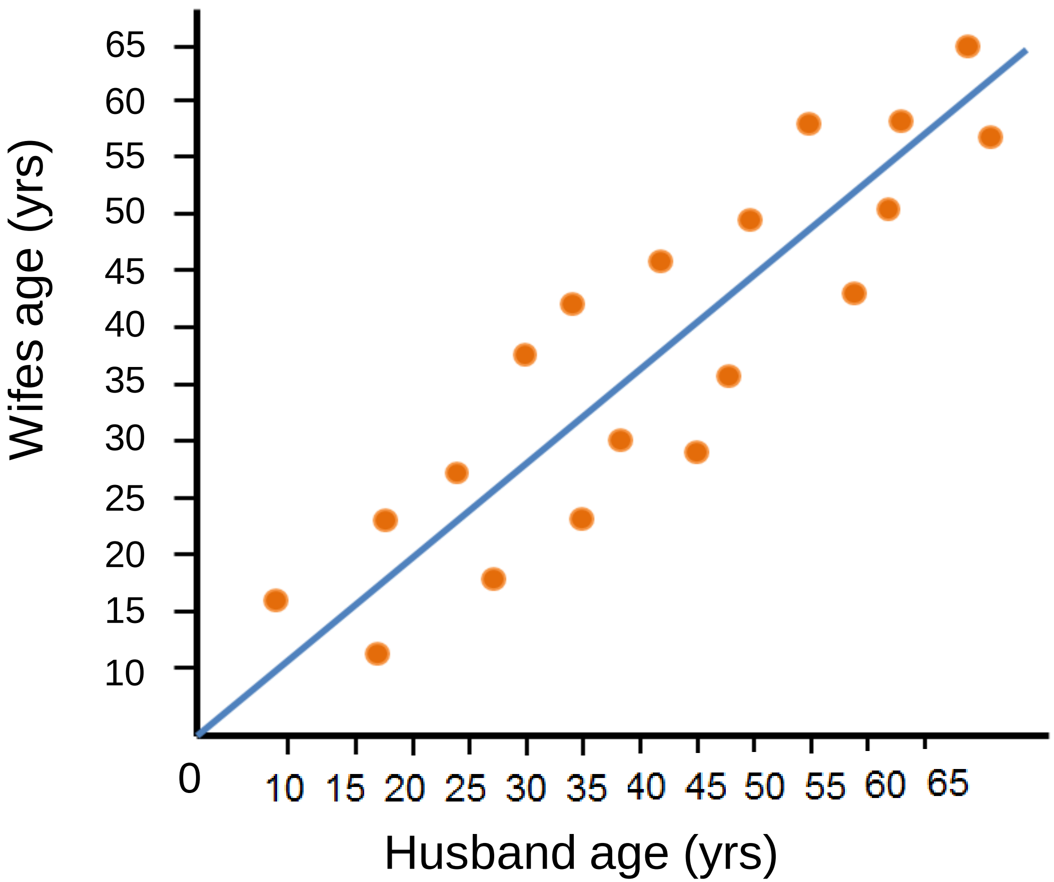


Covariance – to what degree one varies along with the other
Correlation – covariance transformed (-1,1)

Statistical Models

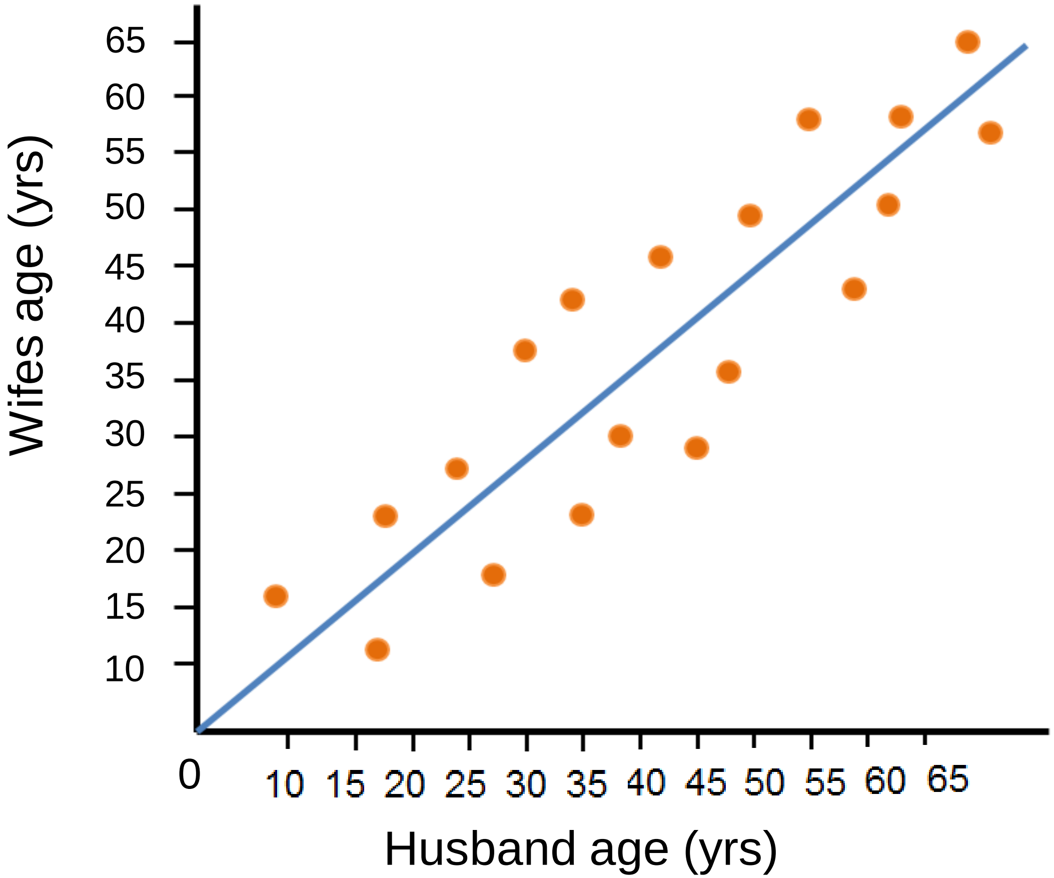
- We observe that husband's and wife's age covary
- What kind of **model** could we fit to this observed data?
- We could theorize that husband's age determines wife's age
  - Older men "choose" older wives
  - This relationship is linear

Linear regression model



Linear regression model

$$WA_i = b_0 + b_1\ HA_i + e_i$$

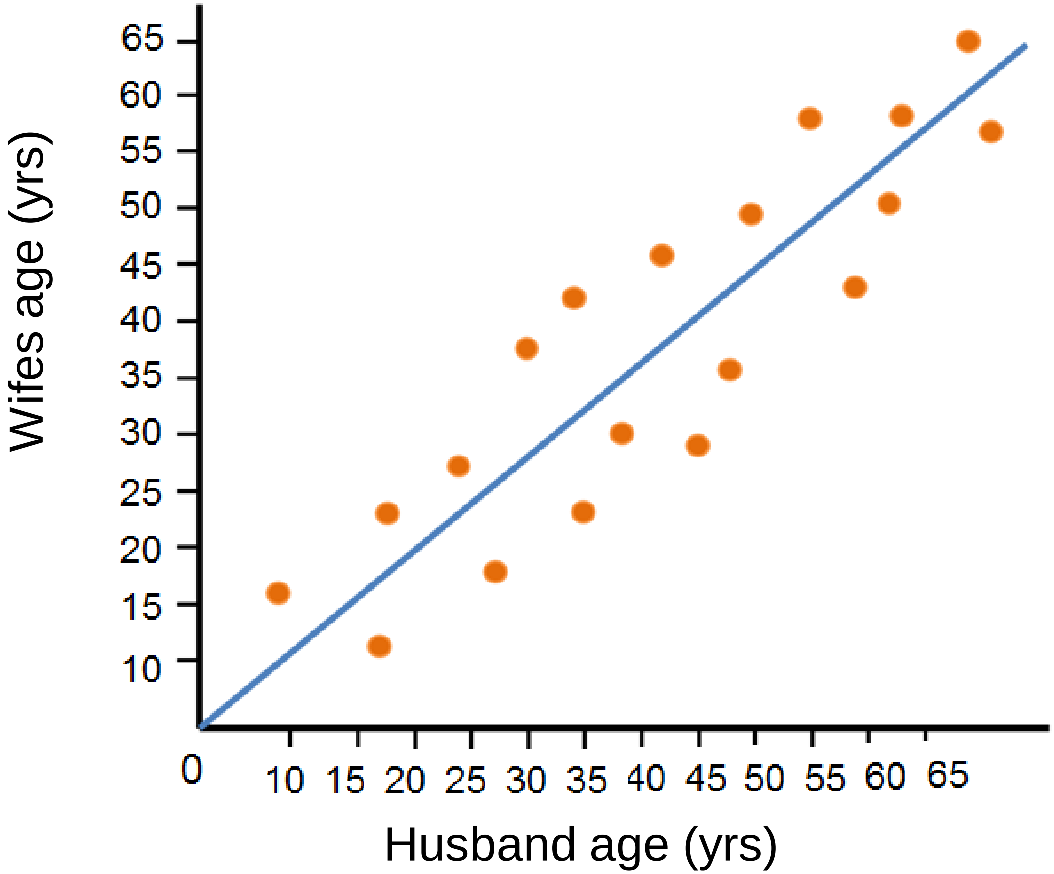


What makes up a model?

- Models are made up of variables and **parameters**
- Parameters are anything in the model that we must **estimate**

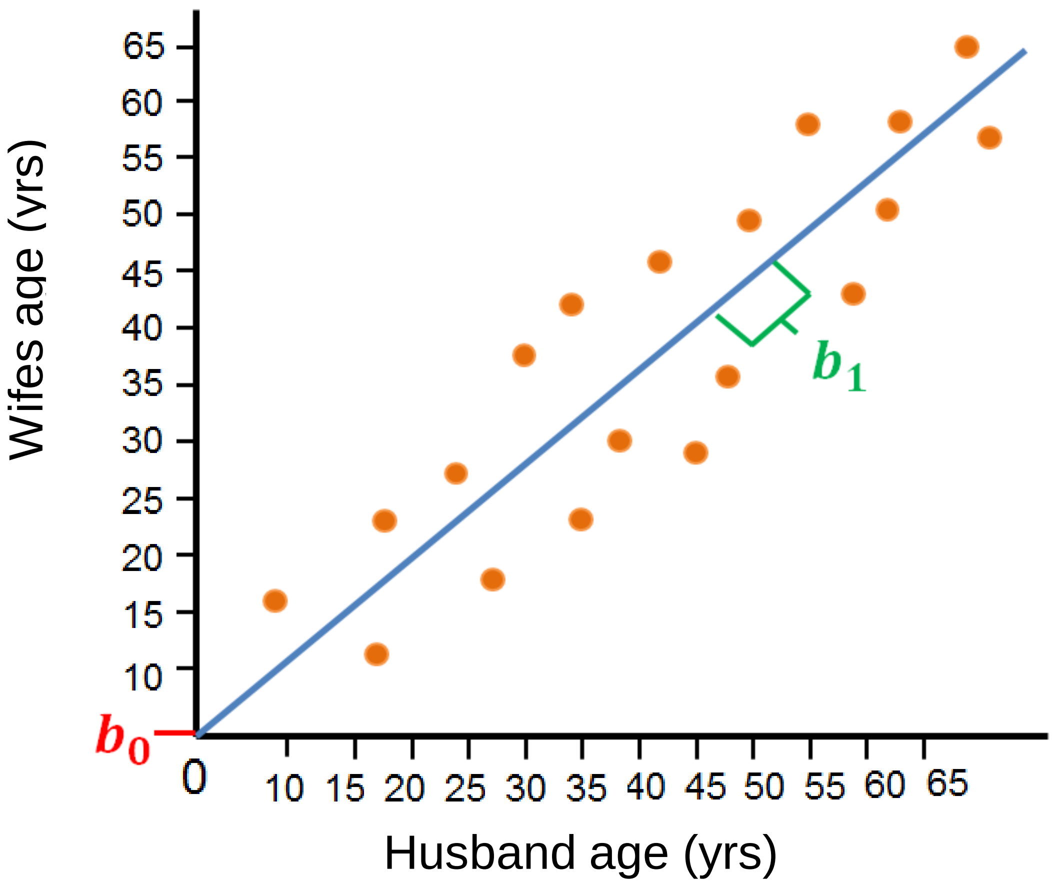
Model Parameters

$$WA_i = b_0 + b_1\ HA_i + e_i$$



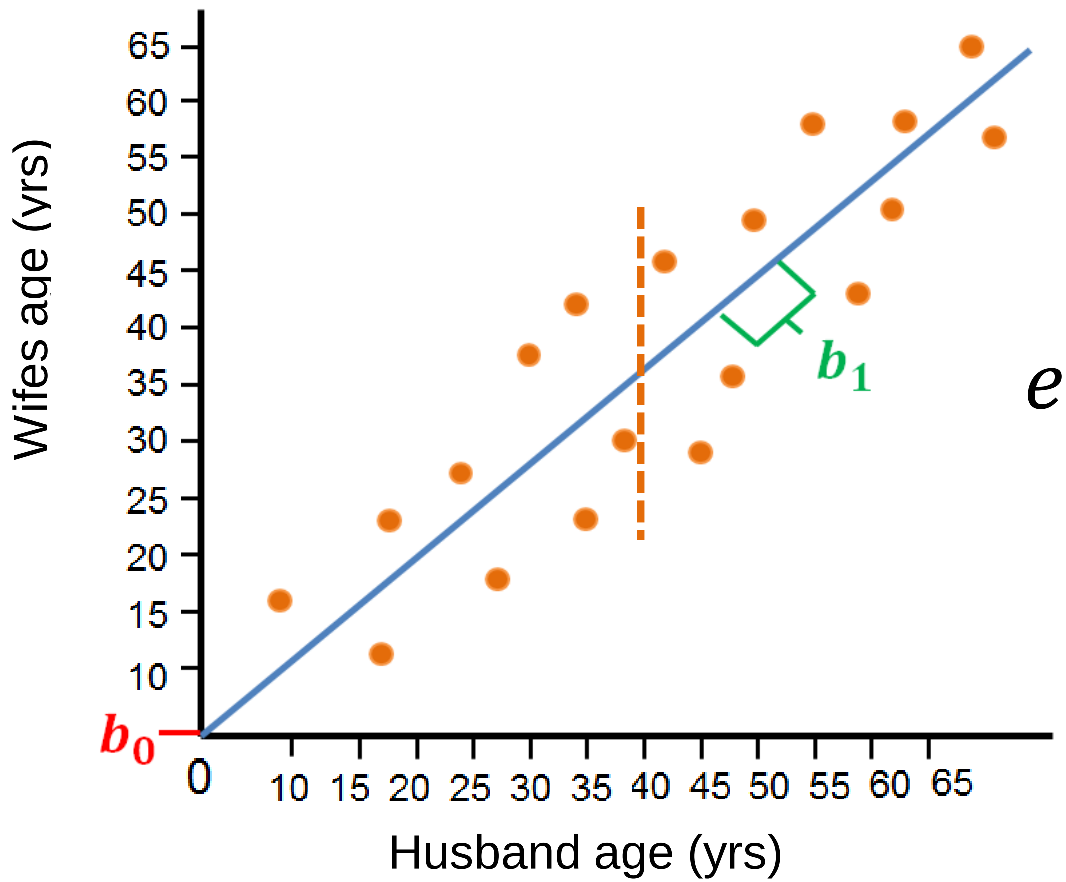
Model Parameters

$$WA_i = b_0 + b_1\ HA_i + e_i$$



Model Parameters

$$WA_i = b_0 + b_1 \, HA_i + e_i$$



$$e_i \sim N\left(0, \sigma_e^2\right)$$

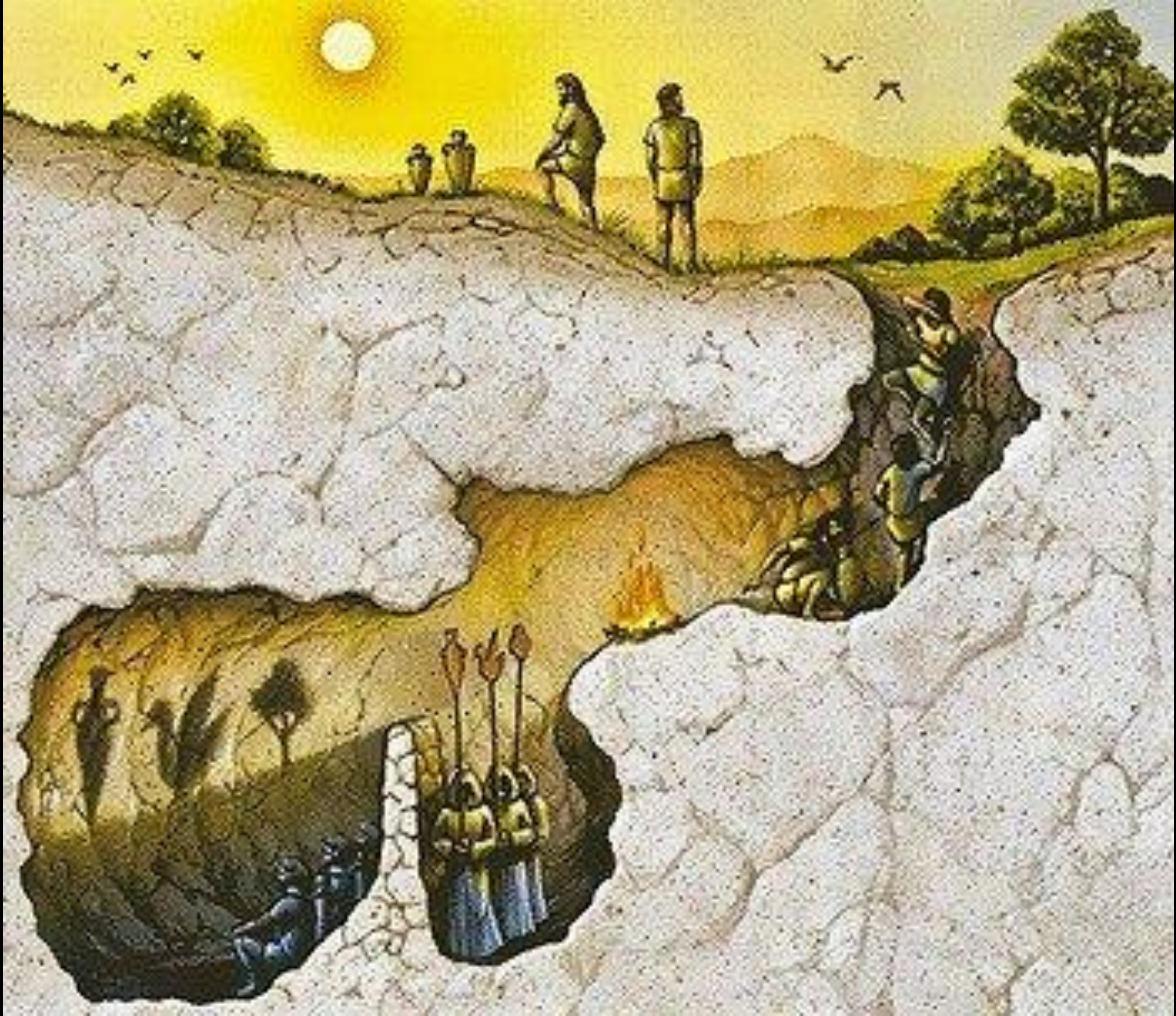
Model Parameters

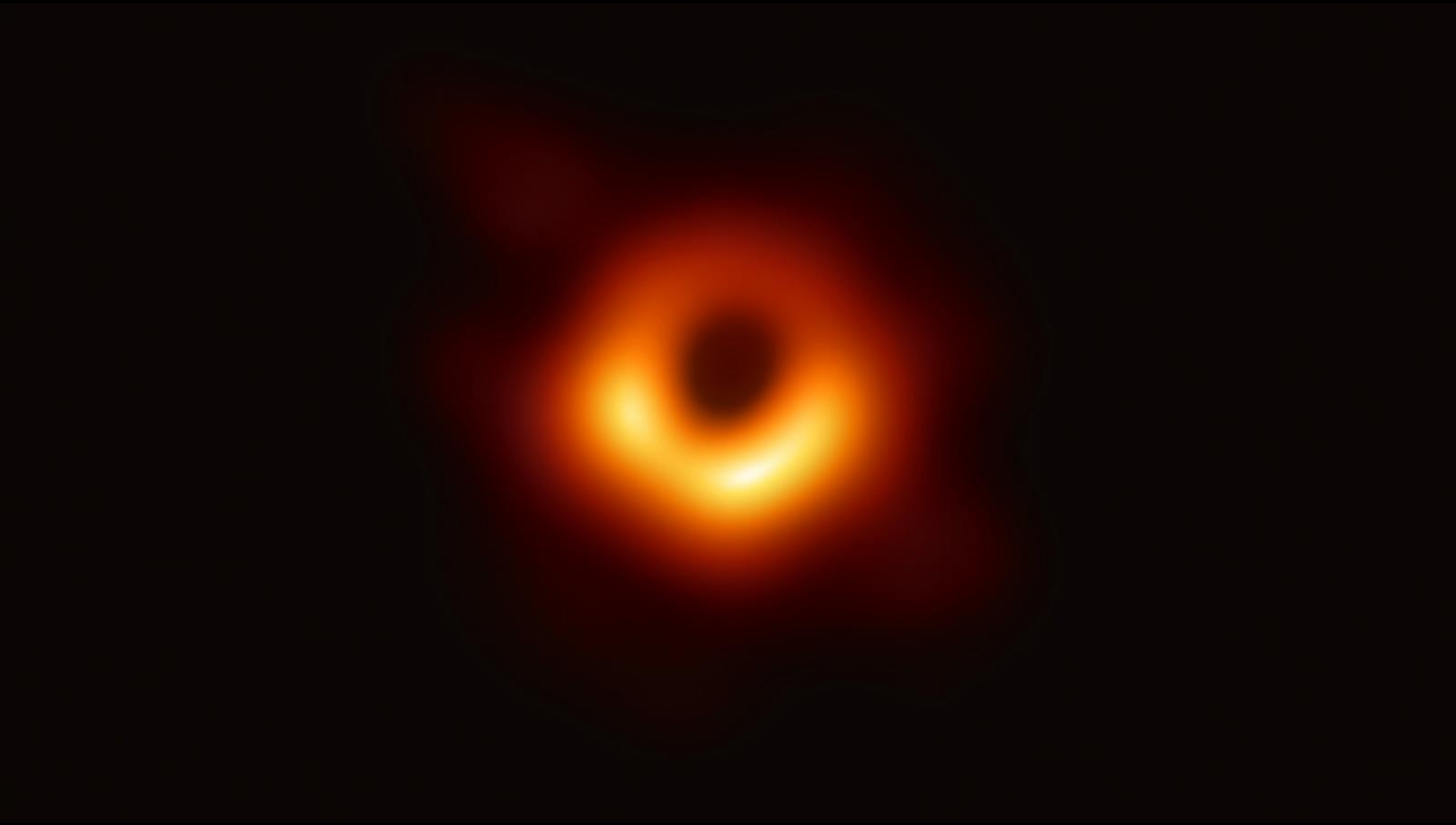
$$WA_i = b_0 + b_1 \, HA_i + e_i$$

Intercept Slope

$$e_i \sim N\,(0, \sigma_e^2)$$

Residual Variance

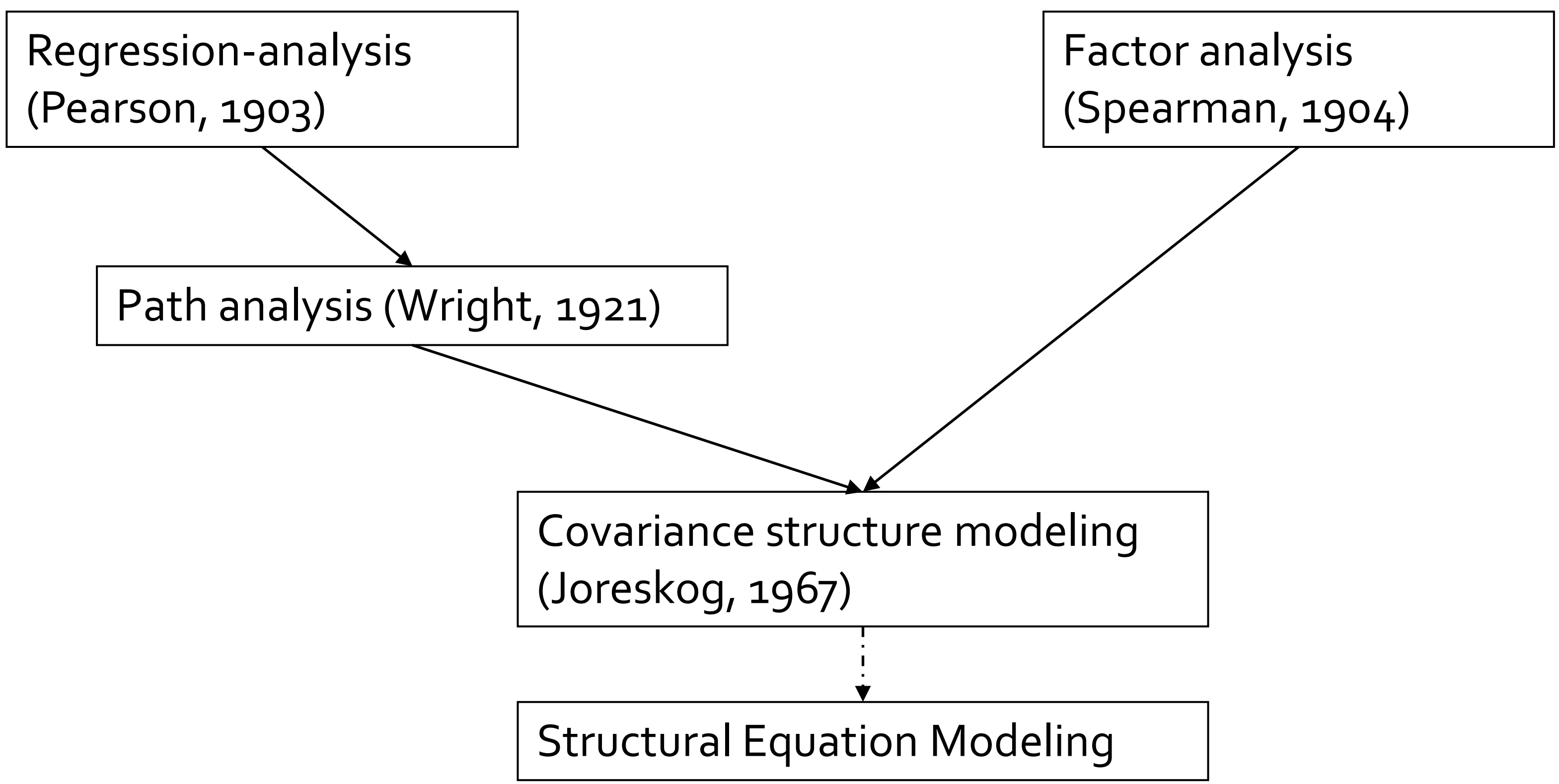




Factor analysis

- Many theories in the social sciences relate to variables which cannot be **directly observed**
  - E.g. depression, personality traits, intelligence
- Instead we try to infer things about these unobservable variables based on what we can observe
  - High scores on IQ test items is taken to reflect high levels of intelligence
- We can call these types of variables **latent variables** or **factors**

History of Structural Equation Modeling

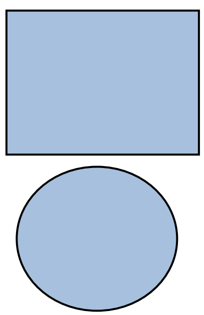


Structural Equation Modeling

- A General Framework encompassing:
  - Linear models: regression, AN(C)OVA, Factor Analysis
    - And any/all combinations
  - Translation of theories with many components
    - Mediation, Moderation
  - Mainly confirmatory but allows for exploratory model search

Path Diagram: Graphical representation of SEM

See the vignette at
https://cjvanlissa.github.io/tidySEM/articles/sem_graph.html



Observed variable

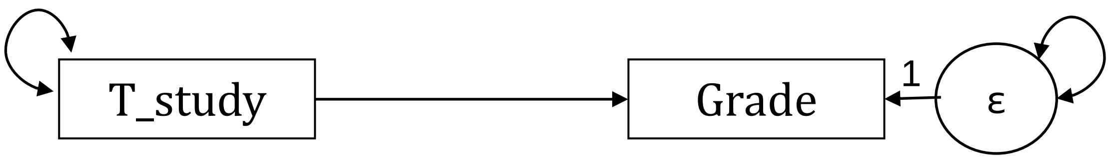
Latent (unmeasured) variable
(or factor)

Regression
- (Theoretical) Causal effect *
- Direct Effect *

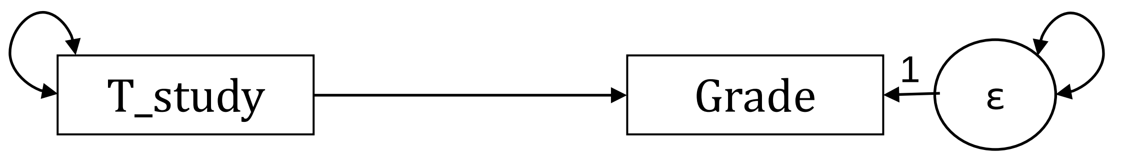
Covariance
- (no causal hypothesis)

Variance
- when going from X to X

Regression model

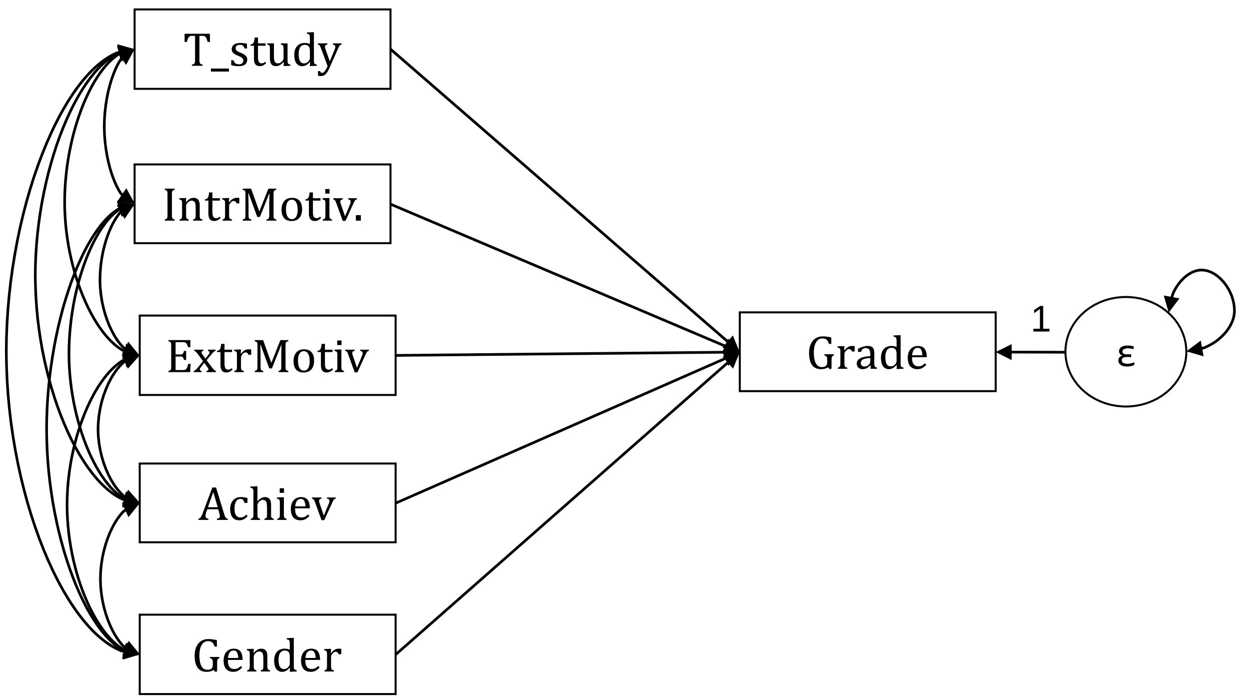


Regression model

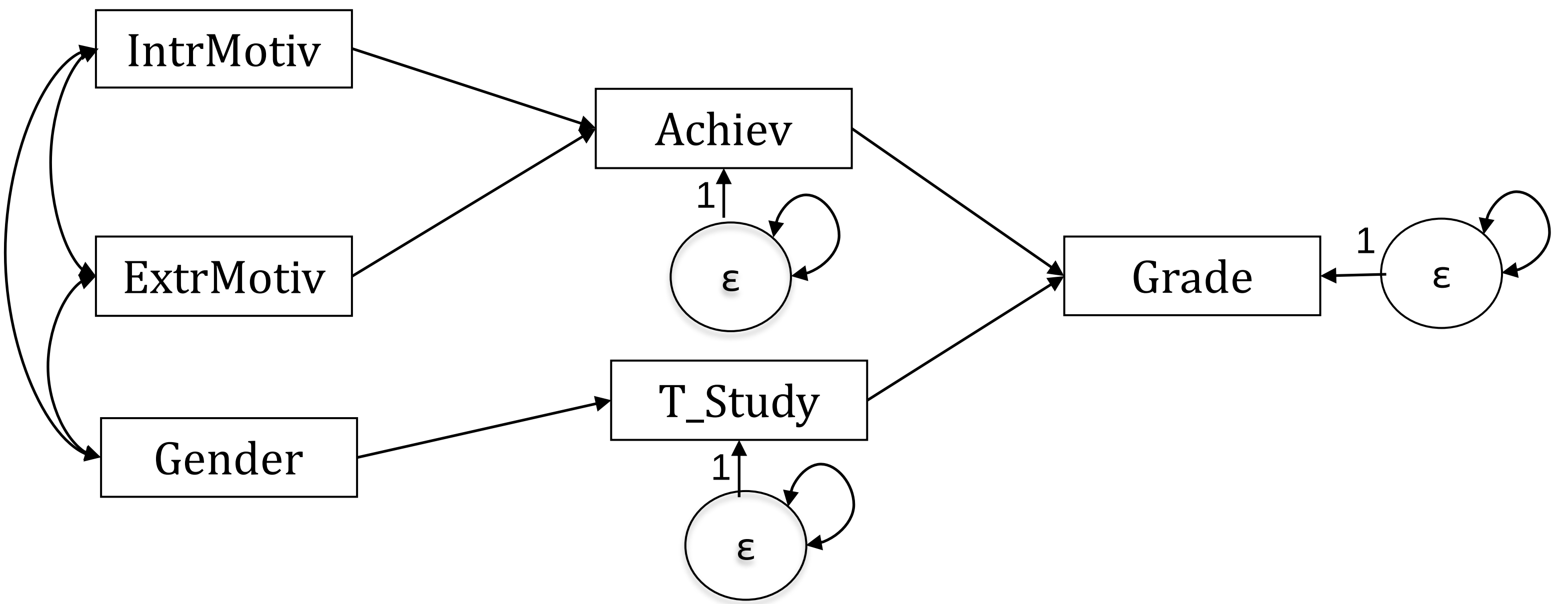


$$Grade_i = b_1 * T\_study_i + e_i$$

Multiple regression model

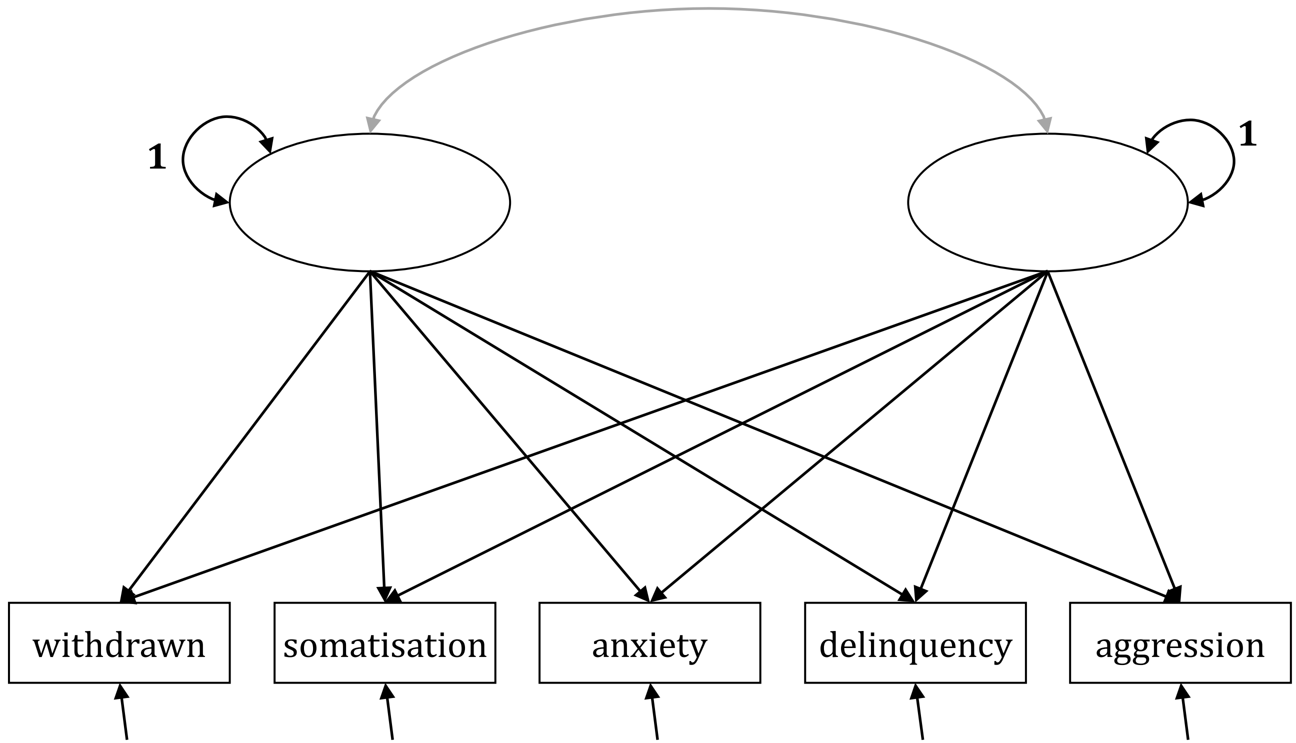


Path model



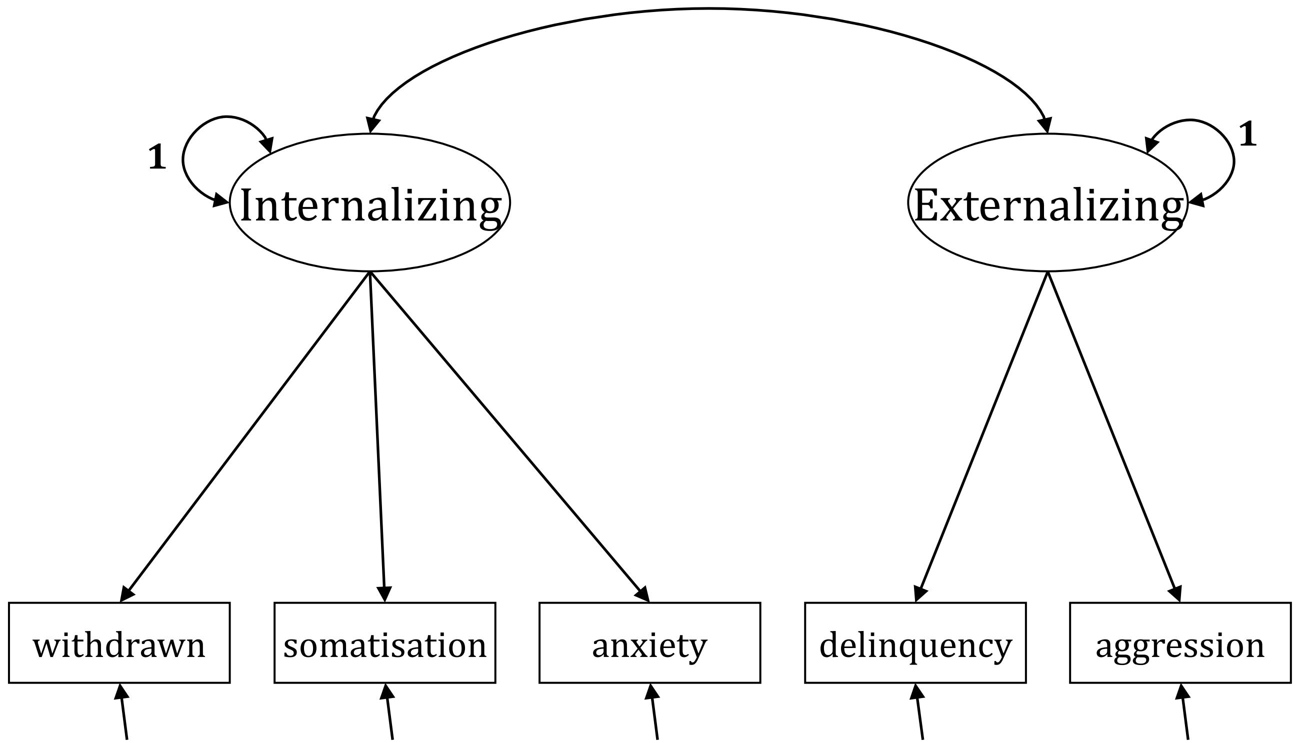
- Direct effects / regression coefficients
- Covariances
- Variances

Exploratory factor analysis model



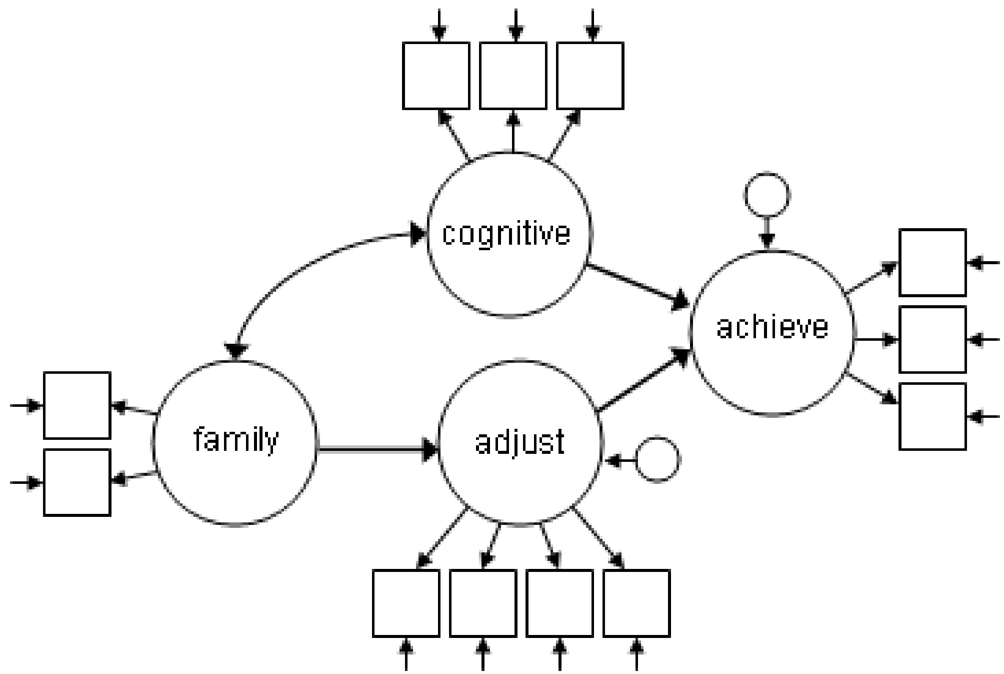
- Factor loadings
- Covariances
- Variances

Confirmatory factor analysis model



- Factor loadings
- Covariances
- Variances

Structural equation model



Interpretation of parameters

- Direct effects, *b*, (X→Y) as regression coefficients
  - If X goes up with 1 point, y is expected to go up with *b* points (controlling for other predictors).
  - If X goes up with 1 SD, y is expected to go up with *b* SD (controlling for other predictors).
- Factor loadings are direct effects from a factor to an indicator
- Covariances (unstandardized) and correlations (standardized)
- Variances and residual variances

Structural Equation Models

- Some assumptions:
  - Multivariate normality of (residuals of) endogenous (outcome) variables (with ML estimation)
    - But there are solutions for categorical data etc (not in this course)
  - Relationships are linear (unless otherwise specified)
  - Independence of observations
  - Exogenous (predictor) variables are measured without error
  - The model is correctly specified

How do Structural Equation Models work?

- They compare an **observed covariance matrix** to a **model-implied covariance matrix**
- Can accommodate complex theories and assumptions
- Evaluate fit: Does the model account for the observed variances and covariances?
  - If our theory says time studying predicts grades, but the observed covariance is zero in our observed data, we have a bad model

FIT AND COMPLEXITY

Choosing Models

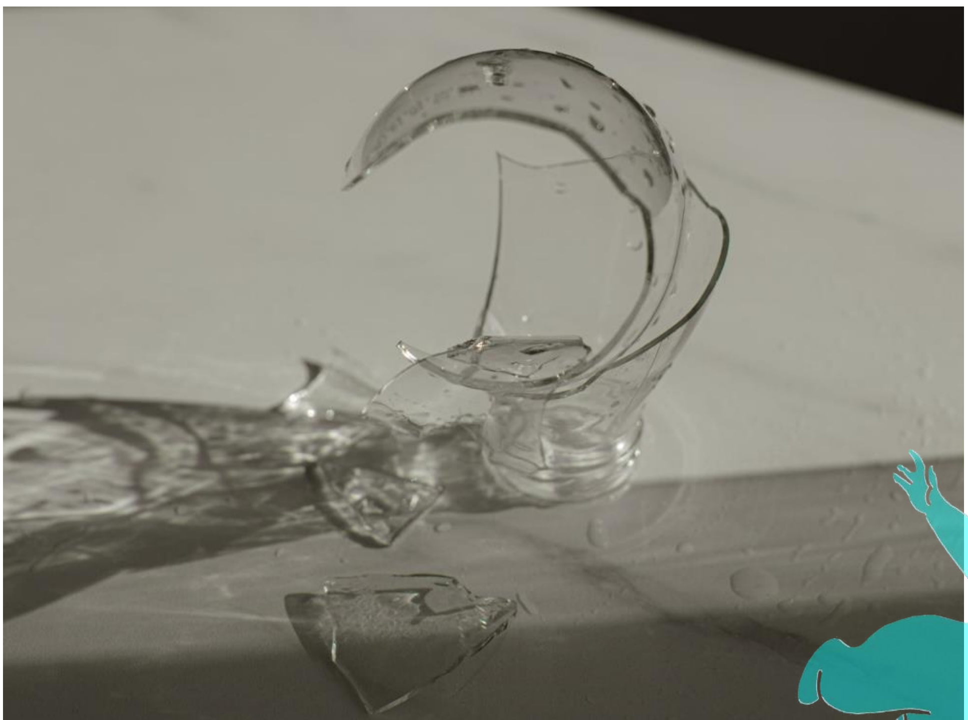
- "All models are wrong but some are useful." George E.P. Box
- "the supreme goal of all theory is to make the irreducible basic elements **as simple and as few** as possible without having to surrender the adequate representation of a single datum of experience." A. Einstein
- "For every complex question there is a simple and wrong solution." H.L Mencken

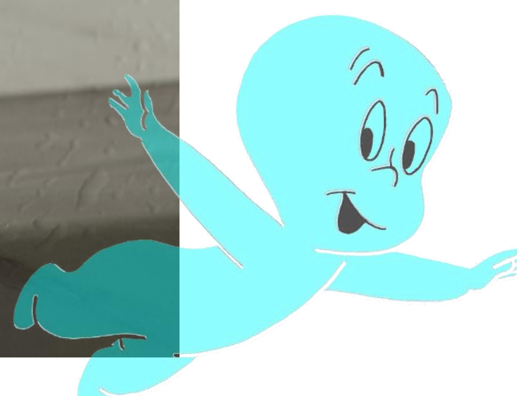
Occam’s Razor

- For each explanation of a phenomenon, there is an extremely large number of possible and more complex alternatives

Occam’s Razor

- For each explanation of a phenomenon, there is an extremely large number of possible and more complex alternatives



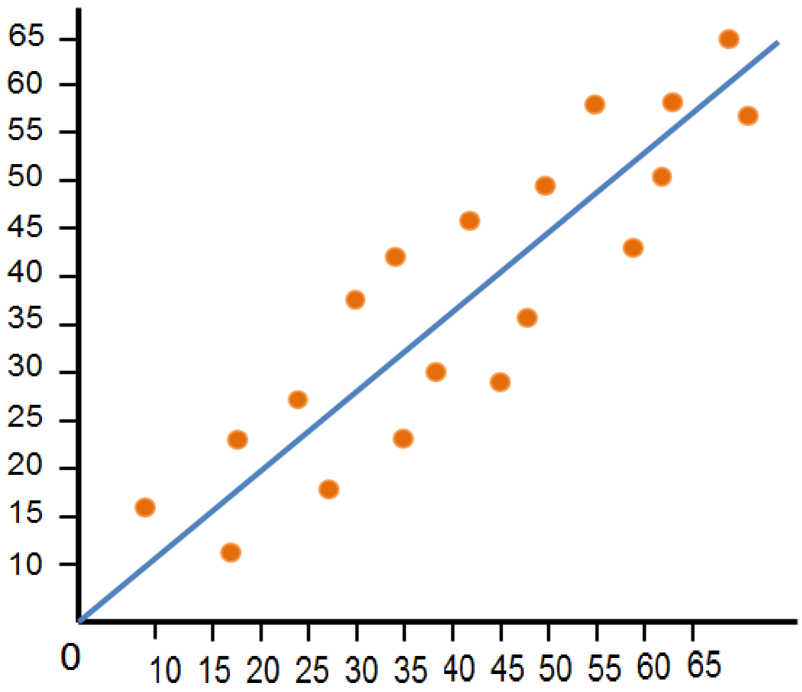


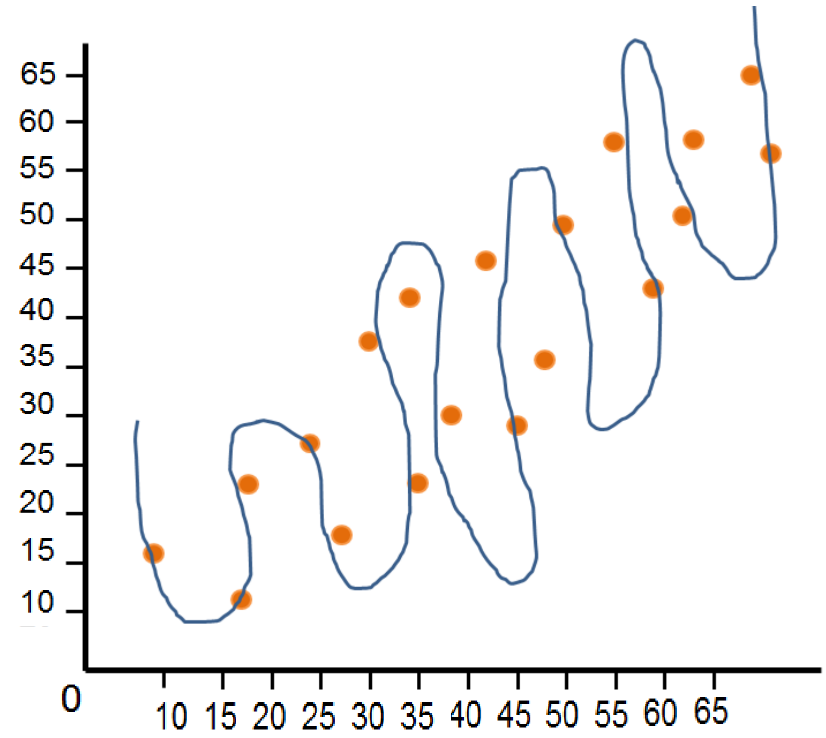
Occam's Razor

- For each explanation of a phenomenon, there is an extremely large number of possible and more complex alternatives

- Prefer simplest possible model for the data *that still fits reasonably well*
  - ***Simple = Parsimonious***

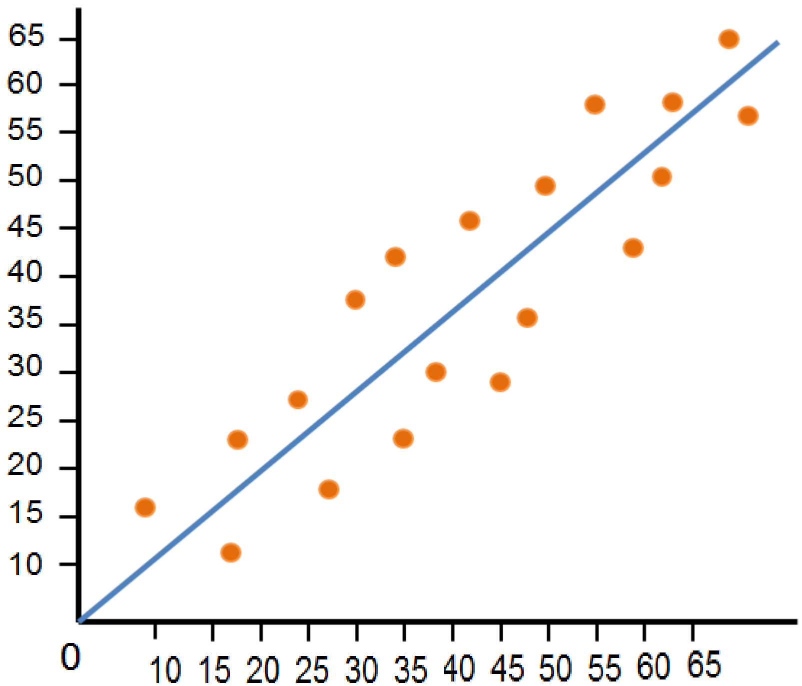
Choosing Statistical Models



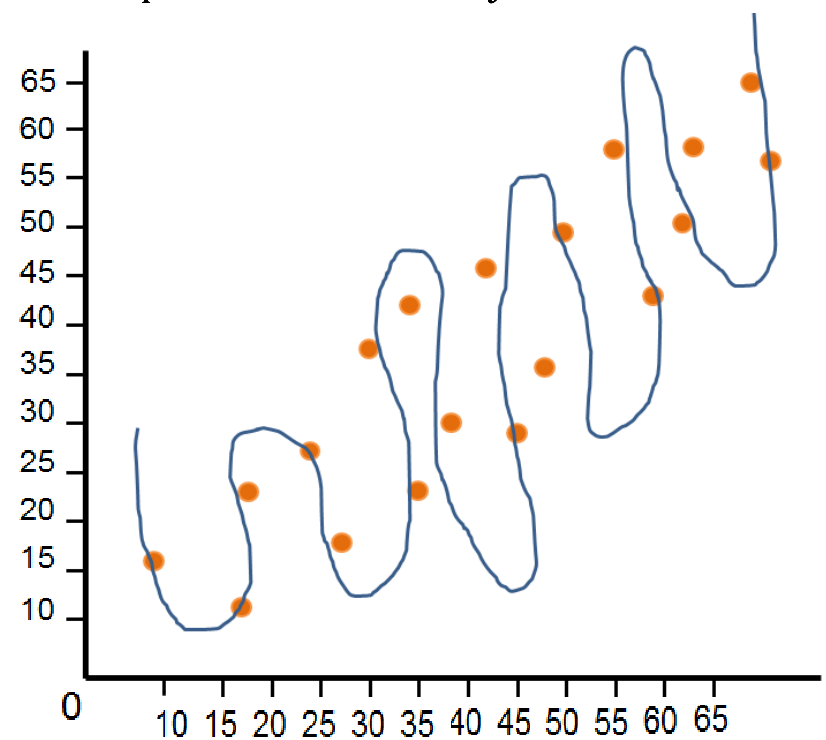


Fit vs complexity

$$WA_i = b_0 + b_1\ HA_i + e_i$$



$$WA_i$$
$$= b_0 + b_1\ HA_i + b_2 HA^2 + b_3 HA^3$$
$$+ b_4 HA^4 + \ \ldots.. + e_i$$



Fit vs Complexity

- Choosing between competing statistical models is a balance between fit and complexity
- **Fit**
  - How well does the model describe the data
- **Complexity**
  - How many parameters are estimated in the model? **

**Other definitions possible – this is sensible when comparing linear models, and so is the definition we will be using throughout

Defining fit

- How well does the model explain the data?
- In regression, the data are individual values on the dependent variable
- In e.g. regression, the data are observations about participants
  - Fit is defined in terms of residual variance in the dependent variable
- In SEM, the data are the **covariance matrix** of your variables

Covariance Matrix

We can summarize relationships between *n_var* variables in a $n\_var \times n\_var$ variance/covariance matrix

	Hus_age	**Wife_age**
Hus_age	s_{Y1}^{2}	
Wife_age	s_{Y1Y2}	s_{Y2}^{2}

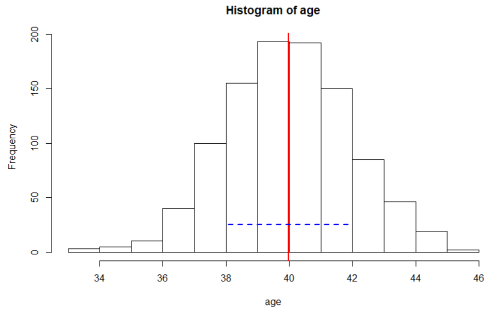
Observed Covariance Matrix

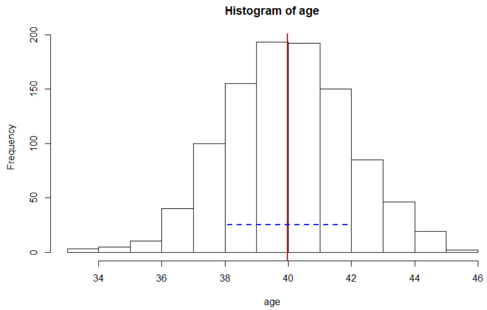
Covariance Matrix

We can summarize relationships between *n_var* variables in a $n\_var \times n\_var$ variance/covariance matrix

	Hus_age	Wife_age
Hus_age	s_{Y1}^2	
Wife_age	s_{Y1Y2}	s_{Y2}^2

Observed Covariance Matrix

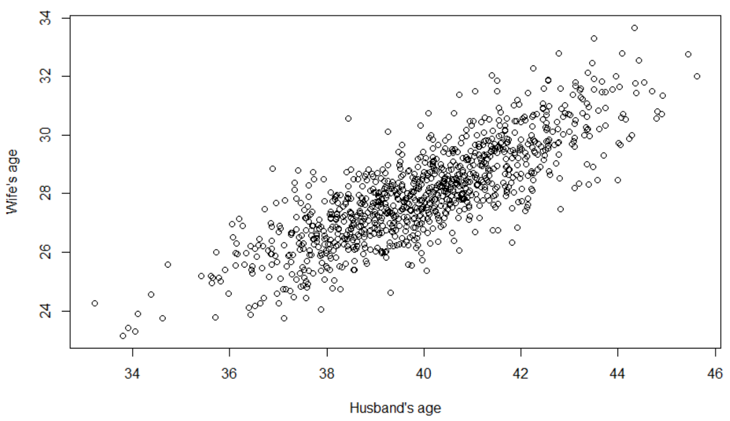




Covariance Matrix

We can summarize relationships between *n_var* variables in a $n\_var \times n\_var$ variance/covariance matrix

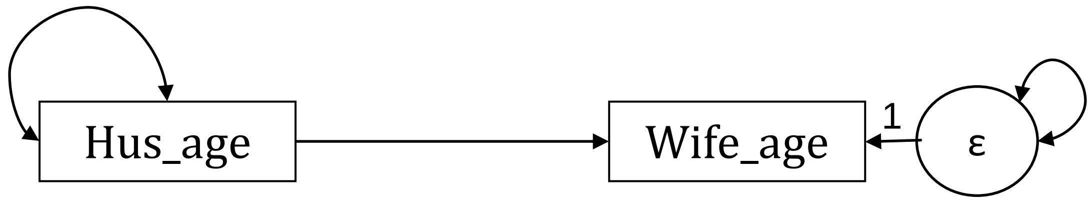
	Hus_age	Wife_age
Hus_age	s_{Y1}^2	
Wife_age	$\boldsymbol{s_{Y1Y2}}$	s_{Y2}^2



Observed Covariance Matrix

Note: We also have information about the ***means*** of each variable.
We will ignore this for now, until **<u>week 4</u>**

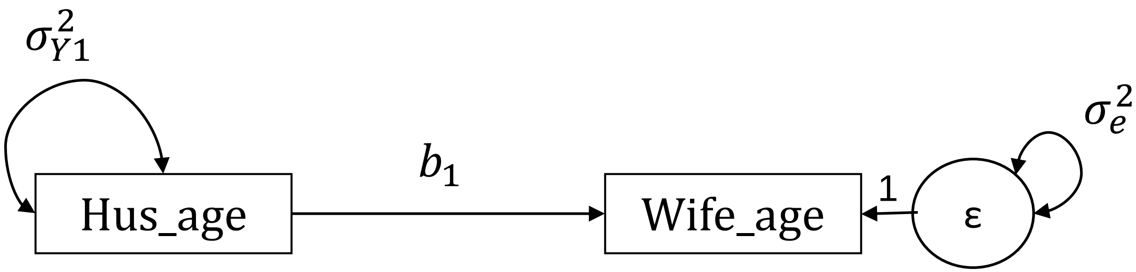
Regression model



	Hus_age	Wife_age
Hus_age	s_{Y1}^2	
Wife_age	s_{Y1Y2}	s_{Y2}^2

Observed Covariance Matrix

Regression model

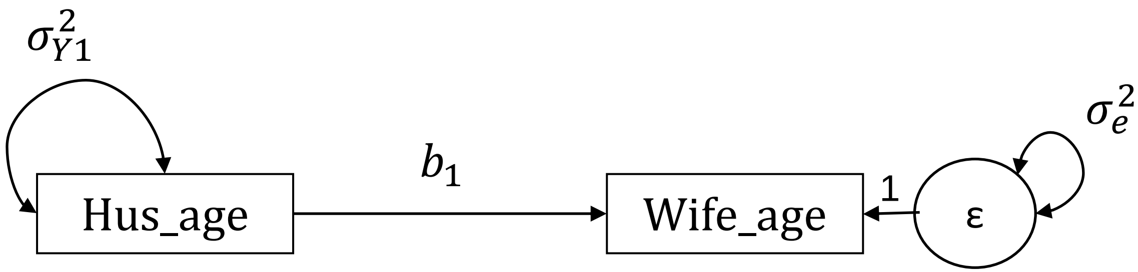


$$Wife\_age_i = b_1 * \; Hus\_age_i + e_i$$

	Hus_age	Wife_age
Hus_age	s^2_{Y1}	
Wife_age	s_{Y1Y2}	s^2_{Y2}

Observed Covariance Matrix

Regression model



$$Wife\_age_i = b_1 * \ Hus\_age_i + e_i$$

	Hus_age	**Wife_age**
Hus_age	σ^2_{Y1}	
Wife_age	$b1$	$\sigma^2_{Y1} b_1 + \sigma^2_e$

Model-implied Covariance Matrix

Defining complexity

The model “explains” the covariances between observed variables.

- Grades and Time-studying co-vary because Time studying has a direct effect on Grades

A good model is:

- Simple (fewest parameters)
- A good description of the data (good fit)

- More degrees of freedom == simpler model (good). But… simpler models fit worse to the data.

Pieces of information

- The “data” in SEM are observed variances/covariances
- These are the pieces of information

	Y1	Y2	Y3	Y4
Y1	s_{Y1}^2			
Y2	s_{Y1Y2}	s_{Y2}^2		
Y3	s_{Y1Y3}	s_{Y2Y3}	s_{Y3}^2	
Y4	s_{Y1Y4}	s_{Y2Y4}	s_{Y3Y4}	s_{Y4}^2

Covariance Matrix

Structural Equation Models

- We can only estimate as many parameters as there are pieces of information
- Estimate parameters to describe the covariance matrix as well as possible
- More variables: more covariances, bigger models

	Y1	Y2	Y3	Y4
Y1	s^2_{Y1}			
Y2	s_{Y1Y2}	s^2_{Y2}		
Y3	s_{Y1Y3}	s_{Y2Y3}	s^2_{Y3}	
Y4	s_{Y1Y4}	s_{Y2Y4}	s_{Y3Y4}	s^2_{Y4}

Covariance Matrix

Degrees of freedom

- We cannot estimate a model with more parameters than pieces of information

- For example, solve for **a**:

  3 = 5 – **a** → a = 2

  b = 5 – **a** → a = ? Impossible to solve

- Our models must be **identified:**
  - Less or equal parameters (q) than observed variances and covariances (p)

Degrees of freedom

- Our models must be **identified:**
  - Less or equal parameters (q) than observed variances and covariances (p)

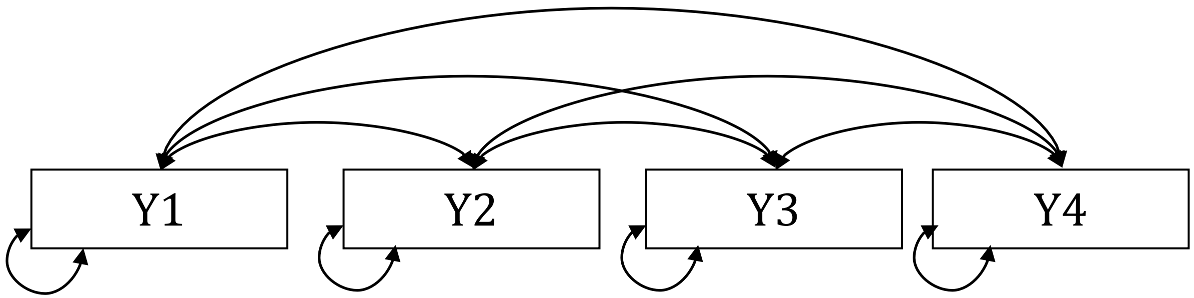
- Degrees of freedom (df) = p – q

- p = nvar*(nvar+1) / 2

	Y1	Y2	Y3	Y4
Y1	4.5			
Y2	2.1	3.9		
Y3	1.9	2.6	4.1	
Y4	2.8	2.5	2.0	4.8

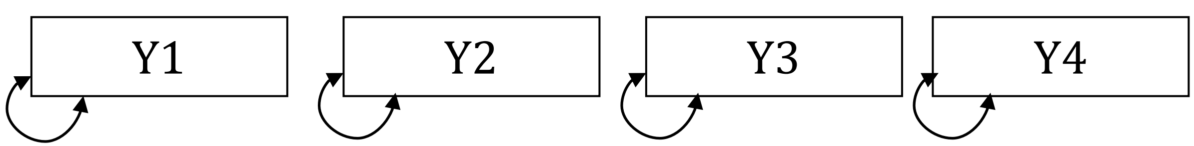
Model complexity in SEM

- Perfectly fitting (but very complex) model:



(Saturated model)

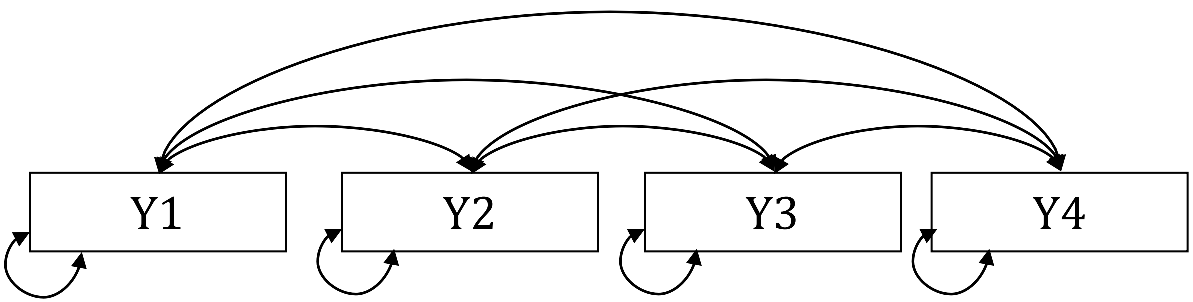
- Very simple (but ill fitting) model:



(Independence model)

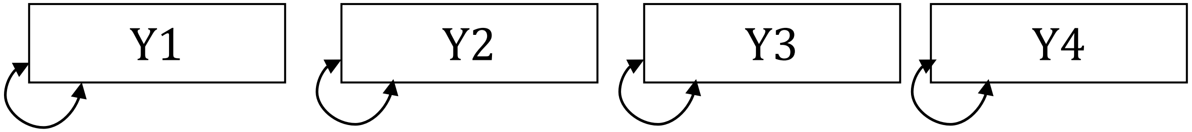
Model complexity in SEM

- Perfectly fitting (but very complex) model:



	Y1	Y2	Y3	Y4
Y1	σ^2_{Y1}			
Y2	σ_{Y1Y2}	σ^2_{Y2}		
Y3	σ_{Y1Y3}	σ_{Y2Y3}	σ^2_{Y3}	
Y4	σ_{Y1Y4}	σ_{Y2Y4}	σ_{Y3Y4}	σ^2_{Y4}

- Very simple (but ill fitting) model:



	Y1	Y2	Y3	Y4
Y1	σ^2_{Y1}			
Y2	0	σ^2_{Y2}		
Y3	0	0	σ^2_{Y3}	
Y4	0	0	0	σ^2_{Y4}

A model for grades

- We observe:
  - IntrMotiv
  - ExtrMotiv
  - Gender
  - Achiev
  - T_study
  - Grades

- How many observed variances-covariances?

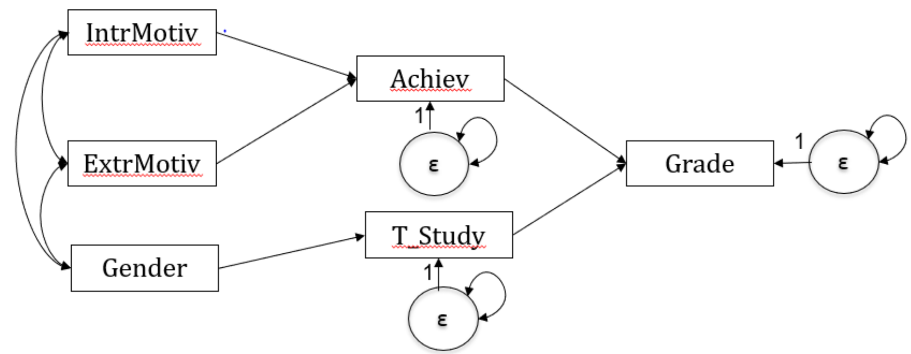
A model for grades

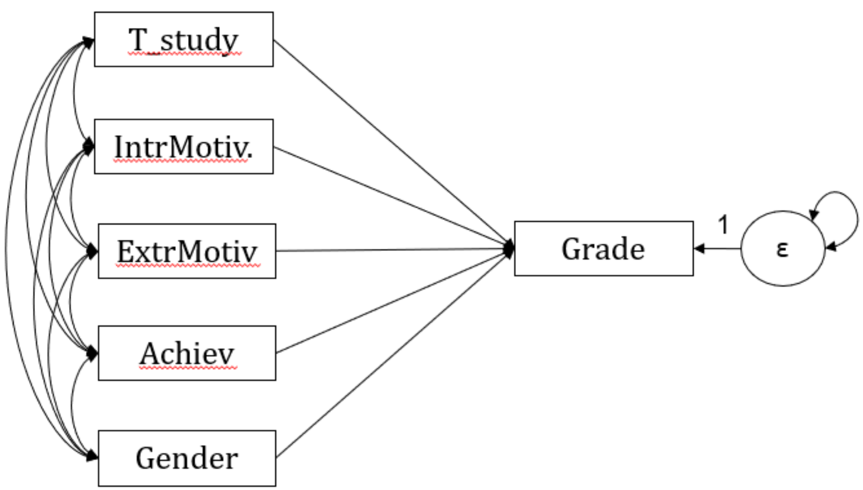
- We observe:
  - IntrMotiv
  - ExtrMotiv
  - Gender
  - Achiev
  - T_study
  - Grades

	Y1	Y2	Y3	Y4	Y5	Y6
Y1	s_{Y1}^2					
Y2	s_{Y1Y2}	s_{Y2}^2				
Y3	s_{Y1Y3}	s_{Y2Y3}	s_{Y3}^2			
Y4	s_{Y1Y4}	s_{Y2Y4}	s_{Y3Y4}	s_{Y4}^2		
Y5	s_{Y1Y5}	s_{Y2Y5}	s_{Y3Y5}	s_{Y4Y5}	s_{Y5}^2	
Y6	s_{Y1Y6}	s_{Y2Y6}	s_{Y3Y6}	s_{Y4Y6}	s_{Y5Y6}	s_{Y6}^2

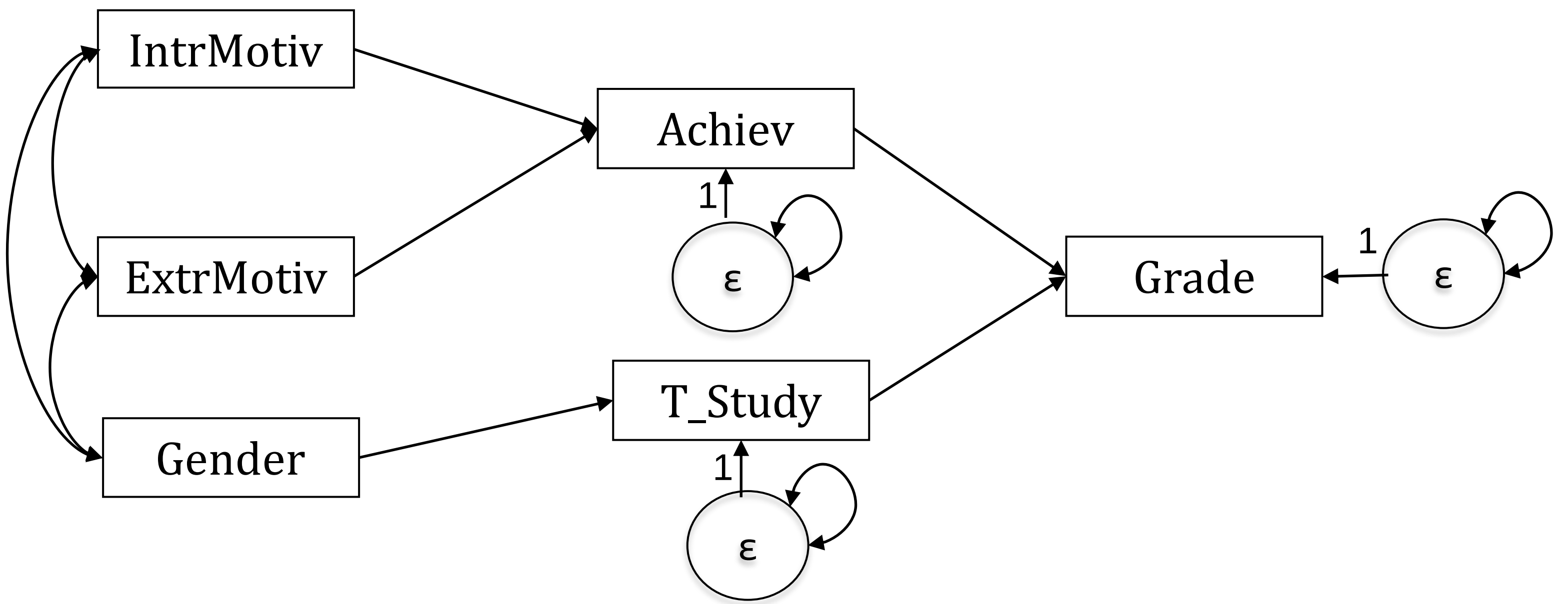
- How many observed variances-covariances? $6 * 7/2 = 21$

Which model is simpler?





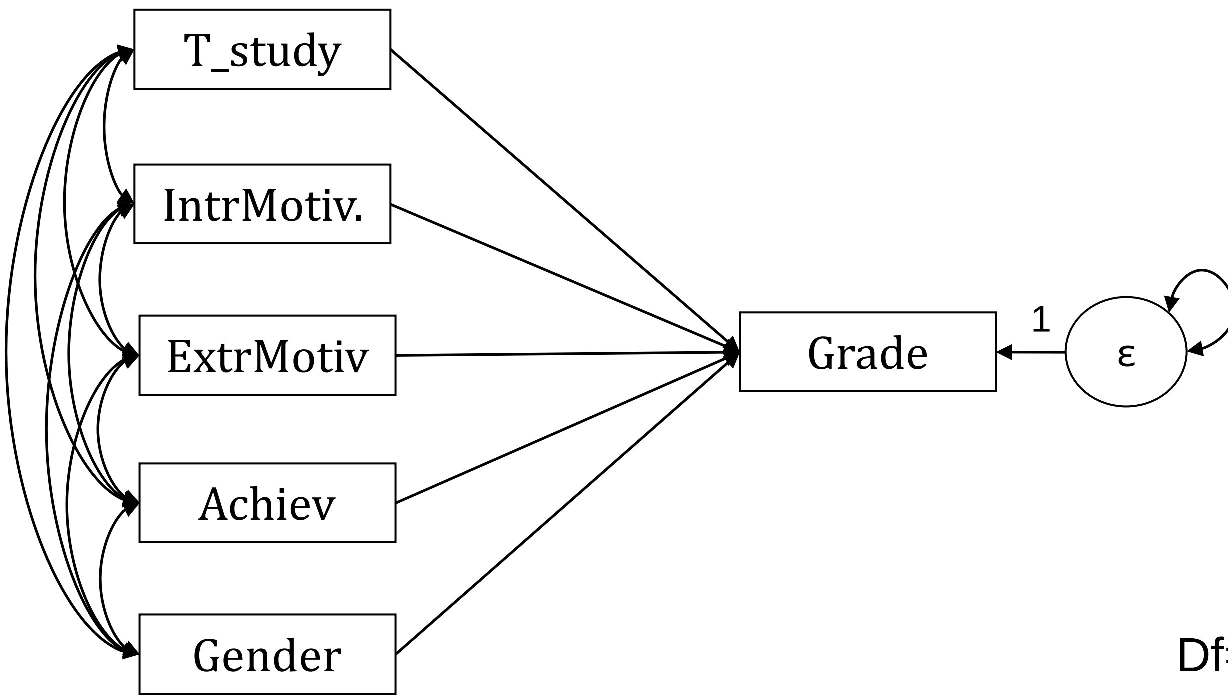
How many degrees of freedom?



Df= 21-14 = 7

3 variances, 3 residual variances
3 covariances, 5 regression coefficients
{14 parameters in total}

Multiple regression model



Df= 21 – 21 = 0

5 variances, 1 residual variances
10 covariances, 5 regression coefficients
{21 parameters in total}

Model fit

- Does the model fit the data? (Exact / approximate fit).
- Yes? Interpret parameter estimates, consider equivalent models. -> **Confirmatory**
- No? Re-specification -> **Exploratory**

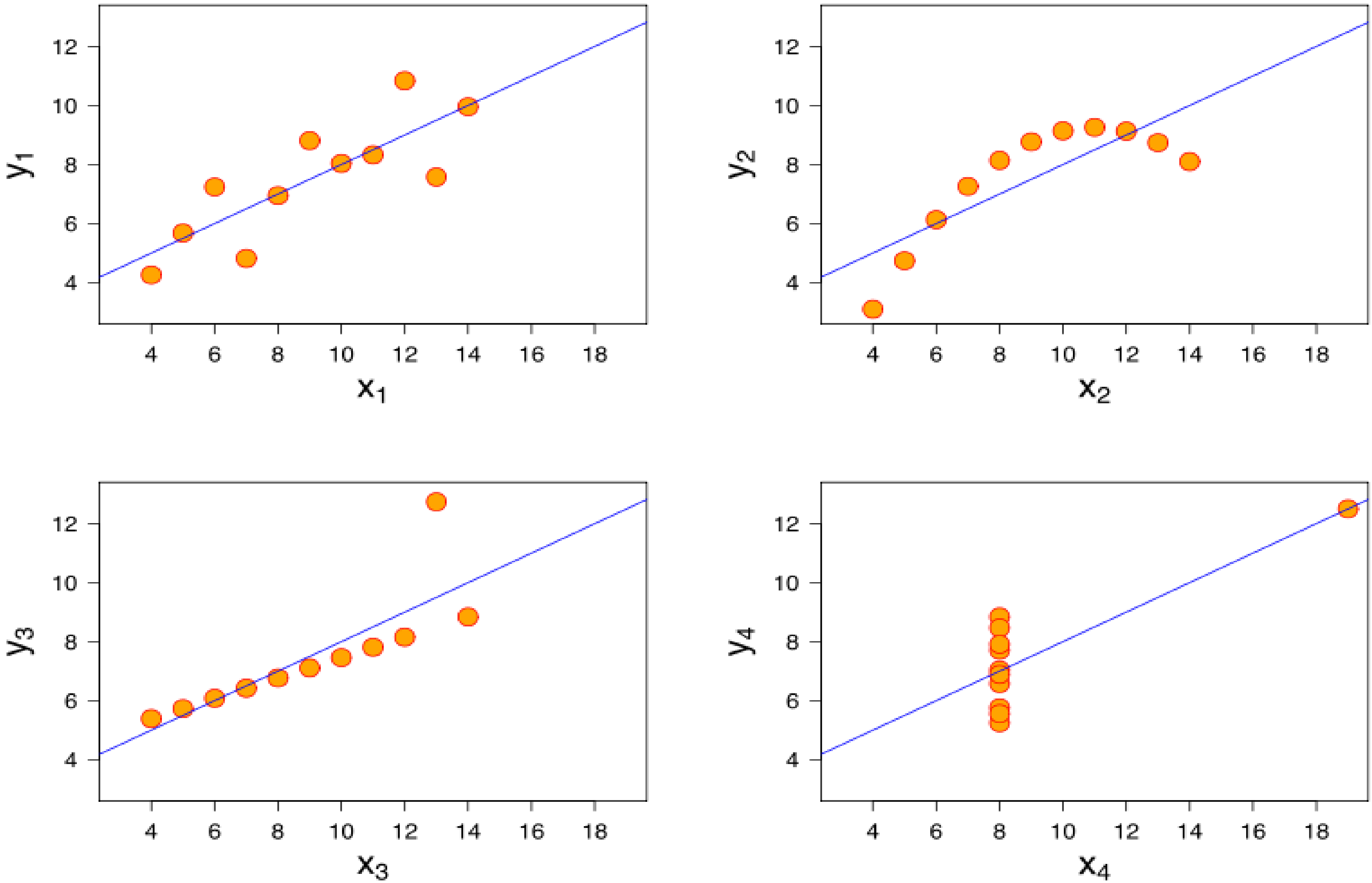
Fit measure	Good	Acceptable	Bad
$X^2_{(df)}$	Non-significant		Significant
RMSEA	< .05	<.08	>.10
CFI	> .95	>.90	

- Many other indices: SRMR, TLI, RNR, GFI, AGFA, AIC, BIC etc. http://davidakenny.net/cm/fit.htm

Model fit: reasons for caution

1. Data can look completely different but have similar covariance matrices

Model fit: reasons for caution



Anscombe's quartet:

$\bar{x} = 9.00$

$\bar{y} = 7.50$

$s_x = 3.16$

$s_y = 1.94$

$r_{xy} = .816$

Model fit: reasons for caution

1. Data can look completely different but have similar covariance matrices
2. Path models can have very different interpretations, but equivalent fits