SEM 2: Structural Equation Modeling

Week 1 - Causal modeling and SEM

Sacha Epskamp

03-05-2018

Course Overview

- Tuesdays: Lecture
- Thursdays: Unstructured practicals
- Three assignments
  - First two 20% of final grade, last 10% of final grade
- Final project
  - Presentations and report, 50% of final grade

Schedule

"Week 1" – Introduction to Structural Equation Modeling

- Thursday May 3 – Lecture + practical
- Tuesday May 8 – Lecture + practical

Week 2 – Causality and equivalent models

- Tuesday May 15 – Lecture + deadline assignment 1
- Thursday May 17 - Practical

Week 3 – Time-series analysis and network models

- Tuesday May 22 – Lecture + deadline assignment 2
- Thursday May 24 - Free time to work on final project

Week 4 – Wrap-up and presentations

- Tuesday May 29 – Recap + practical + (presentations) + deadline assignment 3
- Thursday May 31 - Presentations
- Friday June 1 - deadline final project report

Individual Assignments

Each week, the assignment will be made available 15:00 on Tuesday, and will be due 15:00 the next Tuesday. Each assignment will contribute to 20% or 10% (last week) of your grade.

- Work on the assignments alone.
- Hand in a PDF file and an .R file (in case R was used). If you use Jasp, hand in the Jasp object as well as a screenshot of the options used.
- Make sure your PDF report is as standalone readable as possible. E.g., if you are asked to report a factor loading matrix, then report it in the PDF and not just say "look at .R file".
- Assignments are due before 15:00. If you do not hand in an assignment before 15:00, you will get a 1.
- If you encounter any problems, or have any feedback, please let me know before the deadline, as then I can take it into account or help you.

Final Project

Three options:

1. Perform a SEM analysis on your own data and write a report (individual)
2. Write a manual for semPlot, Onyx, Jasp or Lavaan (individual or with a partner)
3. Research an area or a topic of SEM in more detail and teach fellow students about it

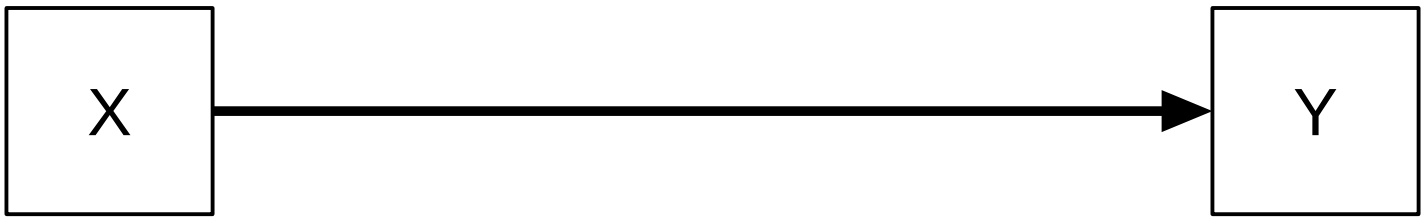
See syllabus on blackboard

- Claim your project using the discussion board on blackboard as soon as possible!
- If you have another idea on a project not listed above, talk to me

Causal modeling

- This course will introduce structural equation modeling (SEM)
- In SEM, we will discuss modeling complex causal hypotheses
- Again, all variables are assumed normally distributed and all associations are assumed linear
- Causal hypotheses can be specified between observed and latent variables
- CFA is a special case of SEM

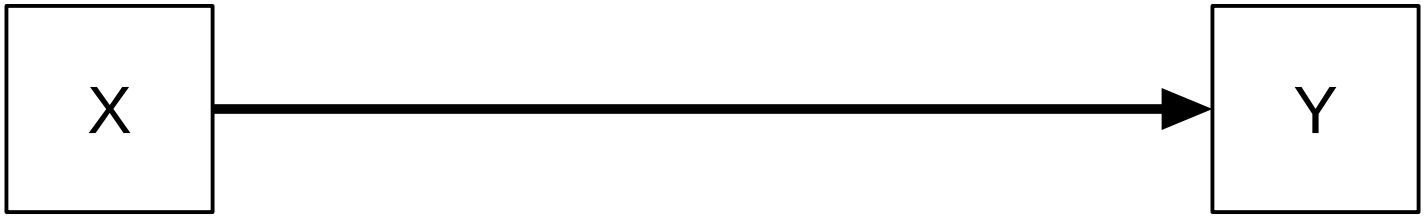
Causal models



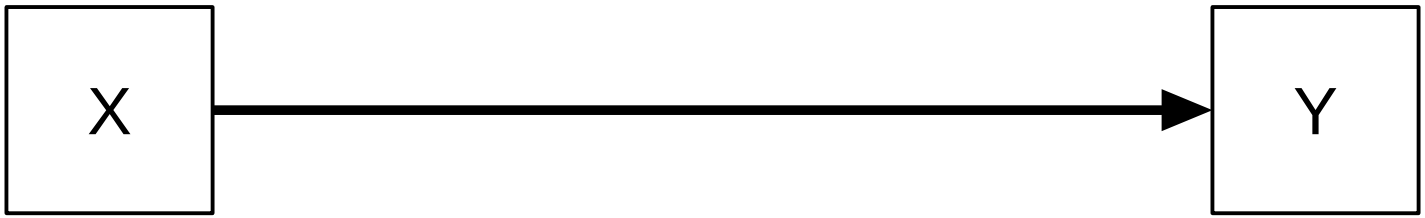
X causes Y

Endogenous and exogenous

- **Exogenous** (independent) variables are variables of which the causal origin are not modeled
  - Exogenous variables have a variance (sometimes not drawn)
  - Exogenous variables, except residuals, are allowed to covary (sometimes not drawn)
  - Latents: ξ (xi); observed: x (x is also used for indicators of latent exogenous variables)
  - Residuals are exogenous
- **Endogenous** (dependent) variables are variables of which the causal origin are modeled
  - Simply stated: endogenous variables have incoming arrows
  - Endogenous variables do not have a variance by themselves
  - Latents: η (eta); observed: y
  - The causal equation for endogenous variables can be derived from the path diagram by summing all incoming edges

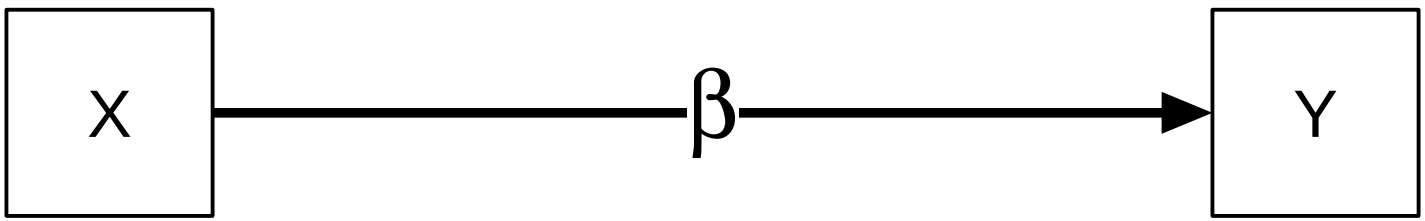


X is exogenous, Y is endogenous

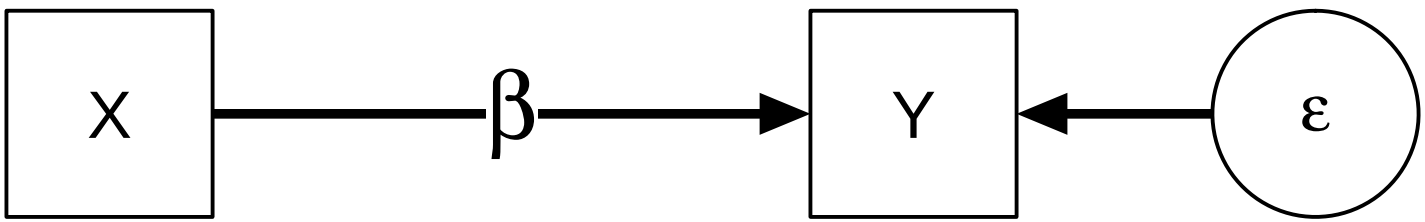


$$y_i = x_i$$

Causal effect goes from right hand side to left hand side. Experimentally changing x will change y, experimentally changing y will not change x

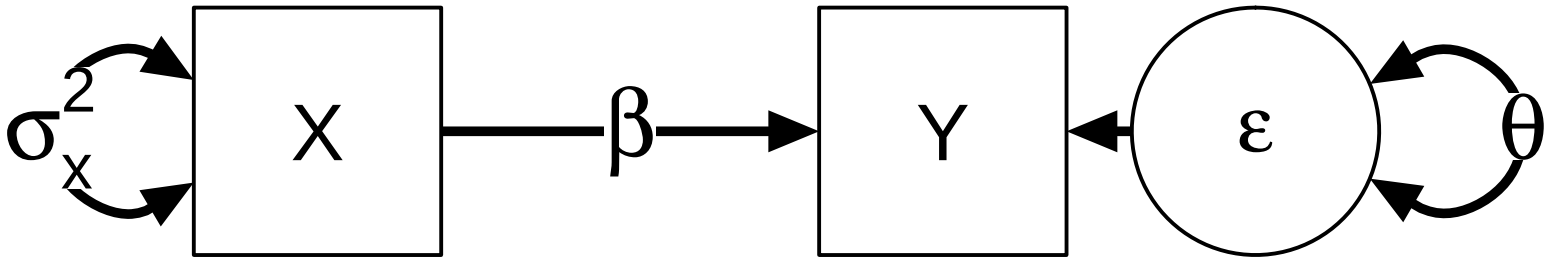


$$y_i = \beta x_i$$



$$y_i = \beta x_i + \varepsilon_i$$

Exogenous variables have a variance (often not drawn)



$$y_i = \beta x_i + \varepsilon_i$$
$$x \sim N(\mu_x, \sigma_x)$$
$$\varepsilon \sim N(0, \theta)$$

$$y_i = \beta x_i + \varepsilon_i$$
$$x \sim N(\mu_x, \sigma_x)$$
$$\varepsilon \sim N(0, \theta)$$

Three observations (variance of x and y and covariance between x and y), three unknowns. Solvable!

$$\mathrm{Var}(x) = \sigma_x^2$$
$$\mathrm{Var}(y) = \beta^2\sigma_x^2 + \theta$$
$$\mathrm{Cov}(x, y) = \beta\sigma_x^2$$

In addition, we can derive that the *expected value* (mean) of y becomes:

$$\mathcal{E}(y) = \beta\mathcal{E}(x).$$

But why?

Expected Values

Suppose someone offers you to play the following game: You throw a dice, if you throw a 6 you win 6 Euro, but if you throw any other number you pay 1 Euro. Would it, statistically, be smart to play this game?

Expected Values

Suppose someone offers you to play the following game: You throw a dice, if you throw a 6 you win 6 Euro, but if you throw any other number you pay 1 Euro. Would it, statistically, be smart to play this game?
Given a 1/6 probability to win 6 Euro and a 5/6 probability to lose 1 Euro, we expect to win:

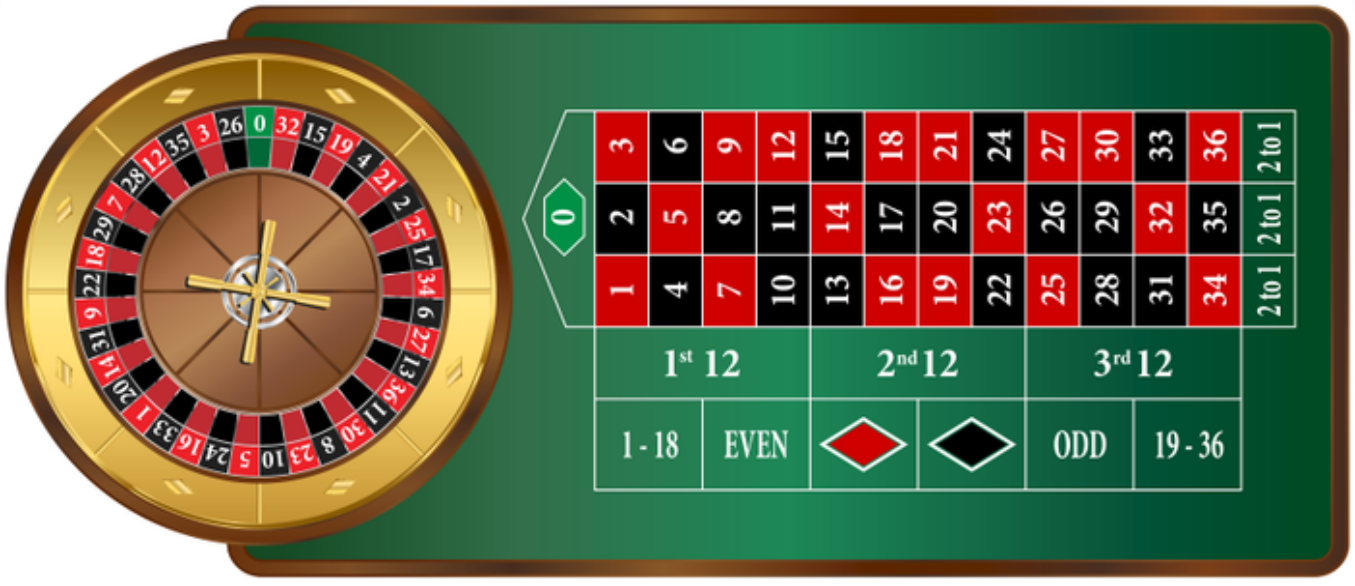
$$6 \times \frac{1}{6} - 1 \times \frac{5}{6} = 0.166\ldots$$

We should play this game!

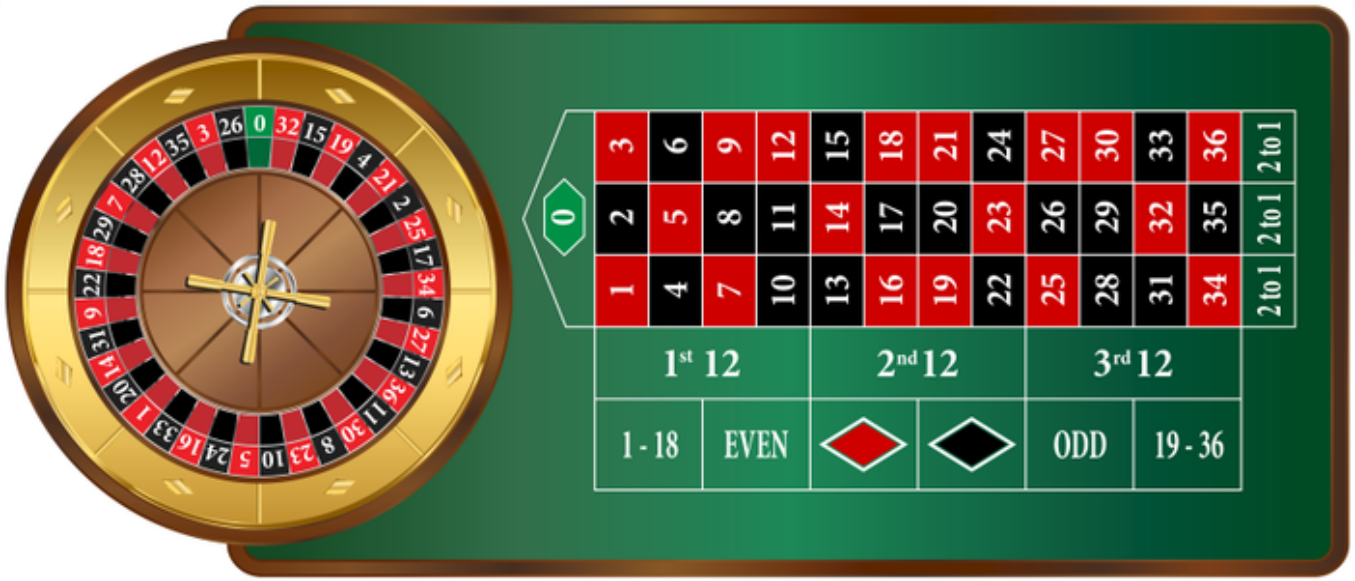
Expected Values

Let $\mathcal{E}(x)$ denote the *expected value* (mean) of a random variable x that can take K different states (e.g, $K = 2$ when you can win or lose). We can obtain the expected value as follows:

$$\mathcal{E}(x) = \sum_{k=1}^{K} x_k \Pr(x = x_k).$$



If you bet on a number and the ball falls on that number, you win 35 times your bet! Should you play this game?



If you bet on a number and the ball falls on that number, you win 35 times your bet! Should you play this game?
Suppose we bet 1 Euro, the expected value is:

$$\frac{1}{37} \times 35 - \frac{36}{37} \times 1 = -0.027$$

So you expect to lose 2.7 cent on average for every bet. This is true for *every possible bet* you can place!

Expected Values

If x is continuous, $k \to \infty$ and $\Pr(x = x_k) \to 0$, but we can still compute the expected value by using the density function $f(x)$ and integration:

$$\mathcal{E}(x) = \int_{\mathbb{R}} x f(x)\, dx.$$

More general, for any function $g(x)$, we can obtain:

$$\mathcal{E}(g(x)) = \int_{\mathbb{R}} g(x) f(x)\, dx.$$

With this we can e.g. proof that if $x \sim N(\mu, \sigma_x^2)$:

$$\mathcal{E}(x) = \mu$$
$$\mathrm{Var}(x) = \mathcal{E}\left((x-\mu)^2\right) = \sigma_x^2$$

For example, see: https://www.statlect.com/probability-distributions/normal-distribution

Expected Values

From:

$$\mathcal{E}(g(x)) = \int_{\mathbb{R}} g(x)f(x)\,dx.$$

we can derive the following rules:

$$\begin{aligned}
\mathcal{E}(\alpha) &= \alpha \\
\mathcal{E}(\alpha x) &= \alpha\mathcal{E}(x) \\
\mathcal{E}(\alpha x + \beta) &= \alpha\mathcal{E}(x) + \beta \\
\mathcal{E}(x + y) &= \mathcal{E}(x) + \mathcal{E}(y) \\
\mathcal{E}(\alpha x + \beta y) &= \alpha\mathcal{E}(x) + \beta\mathcal{E}(y)
\end{aligned}$$

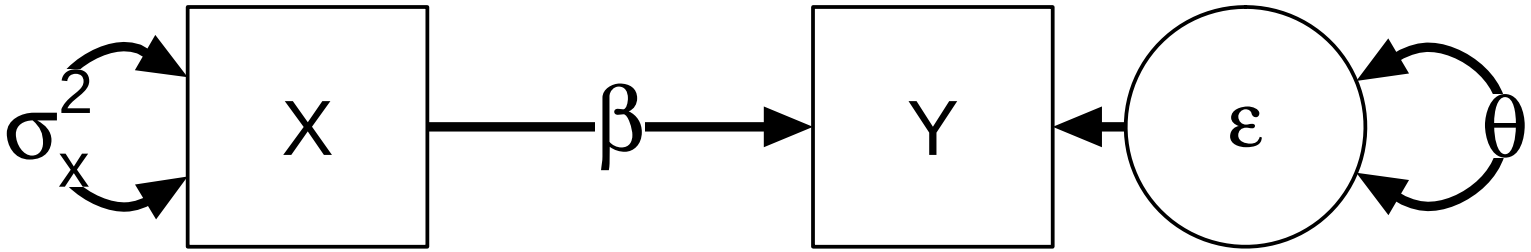
Where α and β are constants (parameter) and x and y are random variables.

Example, the *sample mean* is defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

we can now derive that, if $y \sim N(\mu, \sigma_x^2)$, this is a good estimate for μ:

$$\begin{aligned}
\mathcal{E}(\bar{x}) &= \mathcal{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n}\mathcal{E}\left(\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n} n \mathcal{E}(x) \\
&= \mu
\end{aligned}$$



$$y_i = \beta x_i + \varepsilon_i; x \sim N(\mu_x, \sigma_x); \varepsilon \sim N(0, \theta)$$

We can now derive:

$$\begin{aligned}\mathcal{E}(y) &= \mathcal{E}(\beta x + \varepsilon) \\ &= \mathcal{E}(\beta x) + \mathcal{E}(\varepsilon) \\ &= \beta \mathcal{E}(x)\end{aligned}$$

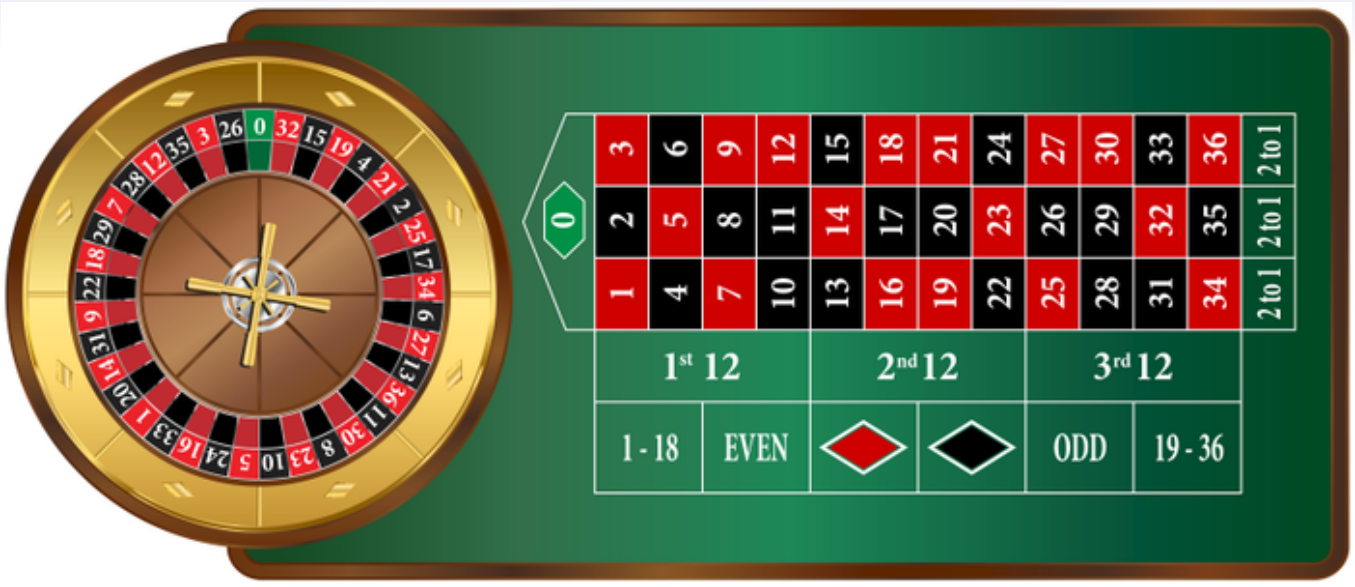
Expected Values

Multivariate generalizations are straightforward:

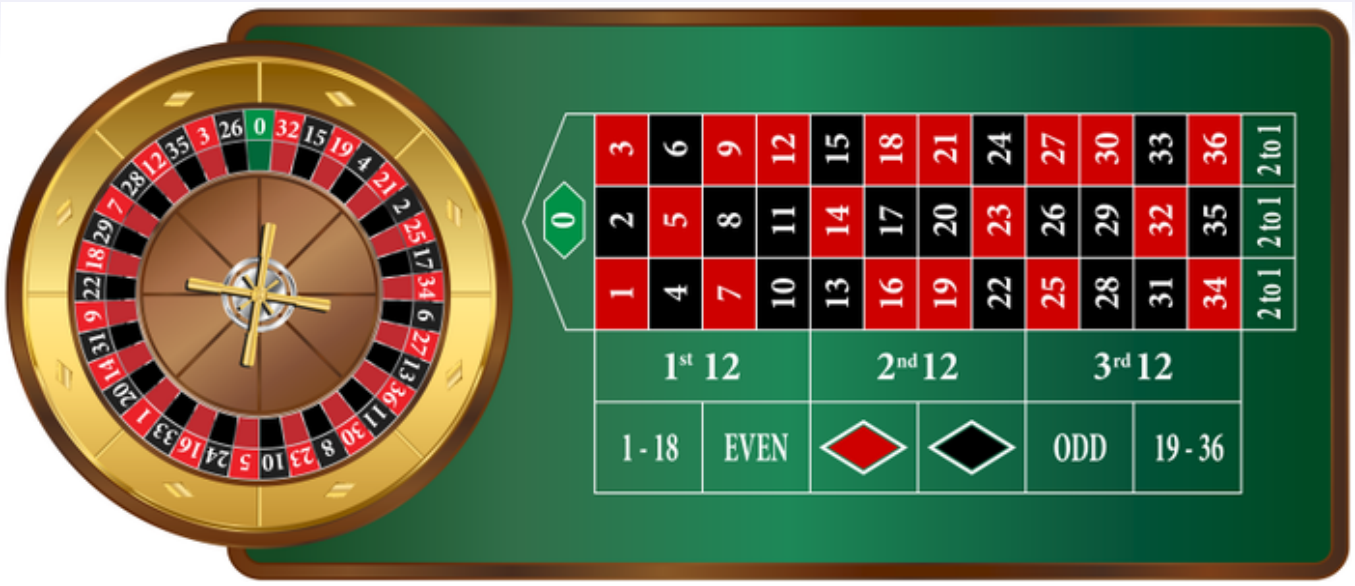
$$\mathcal{E}(\boldsymbol{A}) = \boldsymbol{A}$$
$$\mathcal{E}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\mathcal{E}(\boldsymbol{x})$$
$$\mathcal{E}(\boldsymbol{x}\boldsymbol{B}) = \mathcal{E}(\boldsymbol{x})\boldsymbol{B}$$
$$\mathcal{E}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}) = \boldsymbol{A}\mathcal{E}(\boldsymbol{x}) + \boldsymbol{B}$$
$$\mathcal{E}(\boldsymbol{x} + \boldsymbol{y}) = \mathcal{E}(\boldsymbol{x}) + \mathcal{E}(\boldsymbol{y})$$

Where $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant (parameter) matrices. For example, given $\mathcal{E}(\boldsymbol{\eta}) = \boldsymbol{\alpha}$:

$$\boldsymbol{y}_i = \boldsymbol{\lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i$$
$$\mathcal{E}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{\lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon})$$
$$\mathcal{E}(\boldsymbol{y}) = \boldsymbol{\lambda}\mathcal{E}(\boldsymbol{\eta}) + \mathcal{E}(\boldsymbol{\varepsilon})$$
$$\mathcal{E}(\boldsymbol{y}) = \boldsymbol{\lambda}\boldsymbol{\alpha}$$



Given that $\mathrm{Var}(x) = \mathcal{E}\left((x - \mu_x)^2\right)$ and $\mathcal{E}(x) = \mu_x = -0.027$ for betting 1 Euro in Roulette. What is the variance and standard deviation of our bet?



Given that $\mathrm{Var}(x) = \mathcal{E}\left((x-\mu_x)^2\right)$ and $\mathcal{E}(x) = \mu_x = -0.027$ for betting 1 Euro in Roulette. What is the variance and standard deviation of our bet?

$$\begin{aligned}\mathrm{Var}(\text{bet 1 Euro}) &= \frac{1}{37} \times (35 - -0.027)^2 + \frac{36}{37} \times (1 - -0.027)^2 \\ &= 34.19\end{aligned}$$

$$\mathrm{SD}(\text{bet 1 Euro}) = \sqrt{\mathrm{Var}(\text{bet 1 Euro})} = 5.85$$

Covariance Algebra

Let $\mathrm{Var}(x)$ indicate "the variance of x" and $\mathrm{Cov}(x, y)$ indicate "the covariance between x and y". Given that $\mathrm{Cov}(x, y) = \mathcal{E}\left((x - \mu_x)(y - \mu_y)\right)$, the following rules can be derived:

$$\begin{aligned}
\mathrm{Var}(x) &= \mathrm{Cov}(x, x) \\
\mathrm{Cov}(x, \alpha) &= 0 \\
\mathrm{Cov}(x, y) &= \mathrm{Cov}(y, x) \\
\mathrm{Cov}(\alpha x, \beta y) &= \alpha\beta\mathrm{Cov}(x, y) \\
\mathrm{Cov}(x + y, z) &= \mathrm{Cov}(x, z) + \mathrm{Cov}(y, z)
\end{aligned}$$

Where α and β are constants (parameter) and x, y, and z are random variables.

Covariance Algebra

Some consequences:

$$\begin{aligned}
\mathrm{Cov}(\alpha x + \beta y, z) &= \mathrm{Cov}(\alpha x, z) + \mathrm{Cov}(\beta y, z) \\
&= \alpha \mathrm{Cov}(x, z) + \beta \mathrm{Cov}(y, z) \\
\mathrm{Var}(x + y) &= \mathrm{Var}(x) + \mathrm{Var}(y) + 2\mathrm{Cov}(x, y) \\
\mathrm{Var}(\beta x) &= \beta^2 \mathrm{Var}(x)
\end{aligned}$$

Where α and β are constants (parameter) and x, y, and z are random variables.

Matrix Covariance Algebra

Let $\mathrm{Var}(\boldsymbol{x})$ indicate "the variance–covariance matrix of vector $\boldsymbol{x}$" and $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ indicate "the covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$". Then the following rules can be derived:

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{x}) &= \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{x}) \\
\mathrm{Cov}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{y}) &= \boldsymbol{A}\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{B}^\top \\
\mathrm{Var}(\boldsymbol{B}\boldsymbol{x}) &= \boldsymbol{B}\mathrm{Var}(\boldsymbol{x})\boldsymbol{B}^\top \\
\mathrm{Cov}(\boldsymbol{x} + \boldsymbol{y}, \boldsymbol{z}) &= \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{z}) + \mathrm{Cov}(\boldsymbol{y}, \boldsymbol{z})
\end{aligned}$$

Where $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant (parameter) matrices.

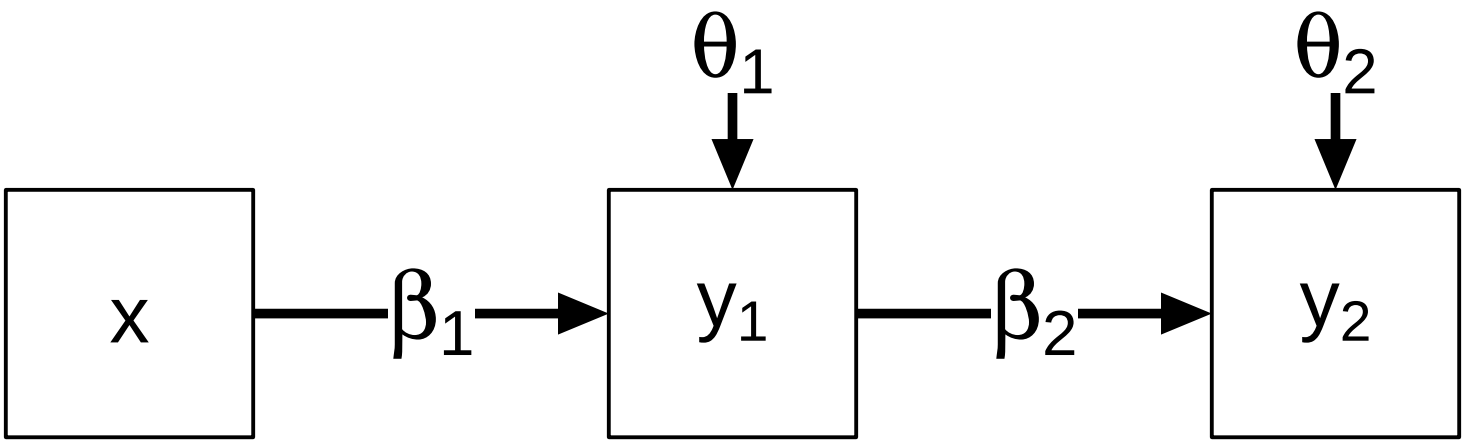
$$
\begin{aligned}
y_i &= \beta x_i + \varepsilon_i \\
\mathrm{Var}(x) &= \sigma_x^2 \\
\mathrm{Var}(y) &= \mathrm{Var}(\beta x + \varepsilon) \\
&= \mathrm{Cov}(\beta x + \varepsilon, \beta x + \varepsilon) \\
&= \mathrm{Cov}(\beta x, \beta x + \varepsilon) + \mathrm{Cov}(\varepsilon, \beta x + \varepsilon) \\
&= \mathrm{Cov}(\beta x, \beta x) + \mathrm{Cov}(\beta x, \varepsilon) + \mathrm{Cov}(\varepsilon, \beta x) + \mathrm{Cov}(\varepsilon, \varepsilon)
\end{aligned}
$$

But since x is not correlated with the residuals, $\mathrm{Cov}(x, \varepsilon) = 0$ and thus:

$$
\begin{aligned}
\mathrm{Var}(y) &= \beta^2 \mathrm{Cov}(x, x) + \mathrm{Cov}(\varepsilon, \varepsilon) \\
&= \beta^2 \mathrm{Var}(x) + \mathrm{Var}(\varepsilon)
\end{aligned}
$$

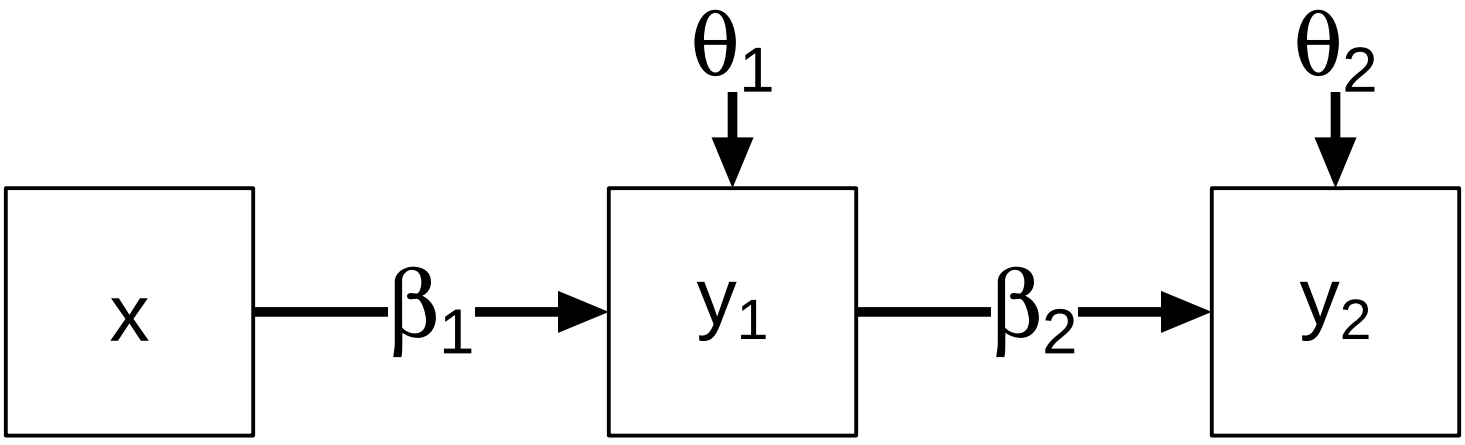
$$
\begin{aligned}
y_i &= \beta x_i + \varepsilon_i \\
\mathrm{Cov}(x, y) &= \mathrm{Cov}(x, \beta x_i + \varepsilon_i) \\
&= \mathrm{Cov}(x, \beta x_i) + \mathrm{Cov}(x, \varepsilon_i) \\
&= \beta \mathrm{Cov}(x, x_i) \\
&= \beta \mathrm{Var}(x)
\end{aligned}
$$

Path analysis

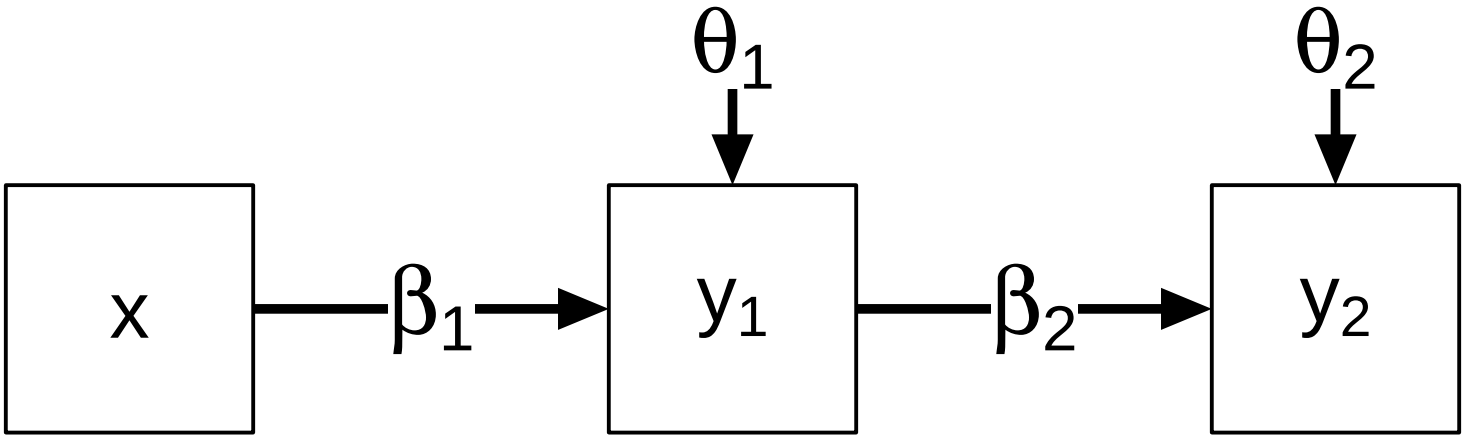


x is exogenous, and both y_1 and y_2 are endogenous. θ_1 is the variance of ε_1. Causal model for y_2:

$$y_{i2} = \beta_2 y_{i1} + \varepsilon_{i2}$$
$$y_{i2} = \beta_2(\beta_1 x_i + \varepsilon_{i1}) + \varepsilon_{i2}$$



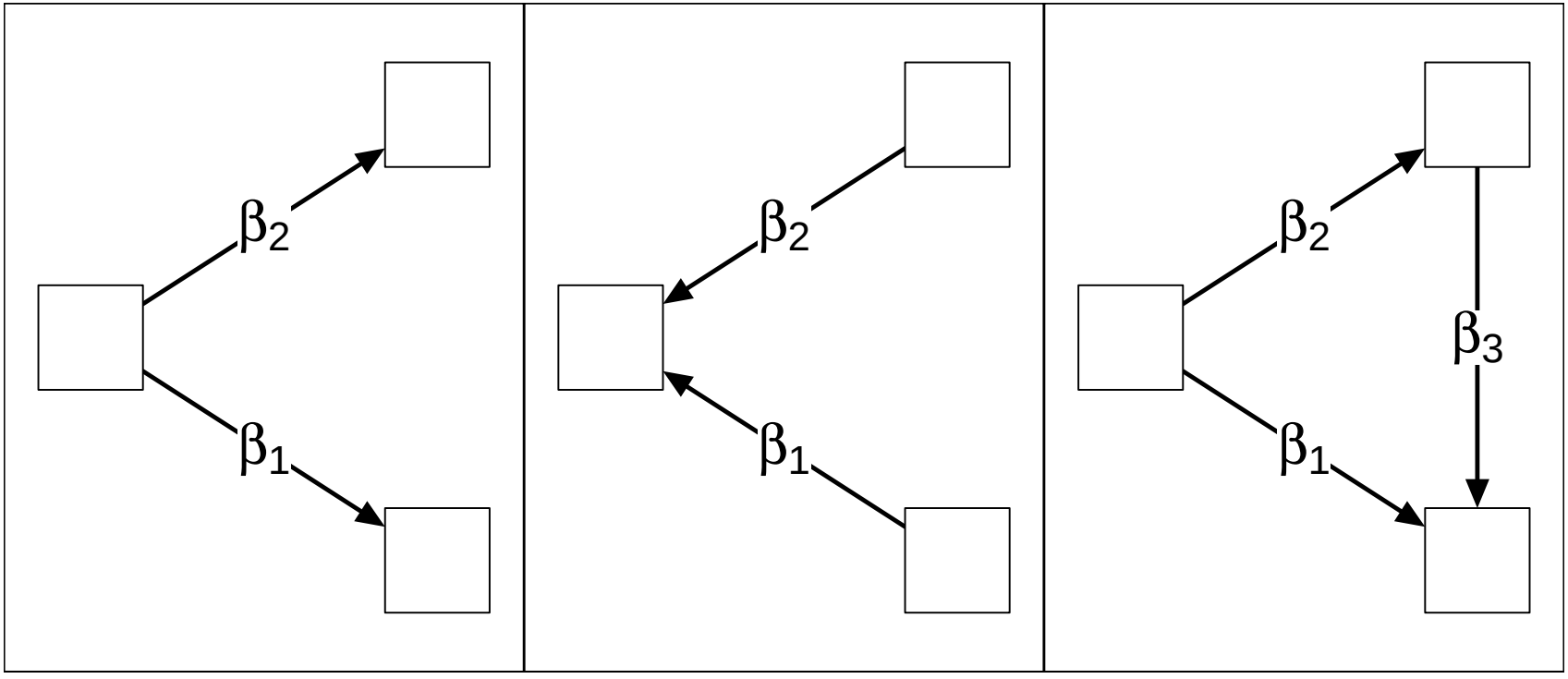
Number of parameters: 2 regressions +2 residual variances +1 exogenous variance (not drawn) = 5, number of observations: 3 variances and 3 covariances. 1 degree of freedom!



Implied covariance between x and y_2:

$$\begin{aligned}
\mathrm{Cov}\,(x, y_2) &= \mathrm{Cov}\,(x, \beta_2(\beta_1 x_i + \varepsilon_{i1}) + \varepsilon_{i2}) \\
&= \mathrm{Cov}\,(x, \beta_2\beta_1 x + \beta_2\varepsilon_1 + \varepsilon_2) \\
&= \mathrm{Cov}\,(x, \beta_2\beta_1 x) + \mathrm{Cov}\,(x, \beta_2\varepsilon_1) + \mathrm{Cov}\,(x, \varepsilon_2) \\
&= \beta_1\beta_2\mathrm{Cov}\,(x, x) \\
&= \beta_1\beta_2\sigma_x
\end{aligned}$$

Practical: identify exogenous (x) and endogenous (y) variables, and derive expressions for expected values and (co)variances of/between all variables, in terms of parameters and expectations/(co)variances of exogenous variables (x and residuals):



Also think about the final project!