# ADVANCED DATA ANALYSIS FOR PSYCHOLOGICAL SCIENCE

## Homework exercises

**Luca Menghini Ph.D.**

luca.menghini@unipd.it

***

Master degree in Developmental and Educational Psychology

University of Padova

2023-2024

# Some instructions to solve the exercises

The present document includes some optional homework exercises on the contents presented during lectures. Similar to the course slides, exercises will be progressively updated as the course progresses.

To check the **exercise solutions**, just look at the `exeRcises.Rmd` file on either Github or Moodle: each exercise text is followed by a chunk of R code showing point-by-point solutions (or some among the many possible solutions).

If you have any doubts on how to solve the exercises, feel free to write me an e-mail or (even better) try writing in the **Moodle forum**, so that all students can see your question and try to reply it. We can also solve some exercise during lectures, but let me know! ;)

# 1. Correlation & regression

For each couple of variables $(x, y)$ generated as specified below:

a) represent univariate (boxplot) and bivariate distributions (scatter plot)

b) compute their correlation

c) use the `lm()` function to get the slope coefficient $\beta_1$ and determinate whether the relationship significantly differs from zero

1. `y <- rnorm(50)` and `x1 <- y`
2. `x2 <- y + 10`
3. `x3 <- rnorm(50)`
4. `x4 <- x3 + 10`
5. Which conclusions can we draw? Which relationship between correlation and regression coefficient?

Source: Pastore, M. (2021). Analisi dei dati in contesti di comunità. Course held at the University of Padova, Academic Year 2021-2022.

# 2. LM assumptions & diagnostics

Using the "*Pregnancy during pandemics*" data* that we saw in class, graphically evaluate the diagnostics of the selected model `m2`:

1. **Linearity**: are model residuals centered on zero?
2. **Normality**: are model residuals normally distributed?
3. **Homoscedasticity**: is residual variance constant over the levels of any predictor?
4. **Independence error-predictor**: are residuals unrelated to any predictor?
5. **Independence of observations**: based on the considered variables (`depr`, `threat`, `NICU`, and `age`), are individual observations independent?
6. **Absence of influential observations**: is there any observation that strongly influence the estimated coefficients?
7. **Absence of multicollinearity**: are predictors mutually unrelated?

*To read the dataset, you can either use the code in `2-multilevel.pdf` slide #10 or download the `pregnancy.RData` file from Moodle/Github ("data" folder) and use the command `load("pregnancy.RData")`

# 3. Towards multilevel modeling

1. Download and read the "*Adolescent insomnia*" dataset `INSA.RData` (Moodle/Github, "data" folder)
2. Explore the variables `dayNr` (day of assessment), `stress` (bedtime rating of daily stress), `insomnia` (categorical: insomnia vs. controls), and `TST` (total sleep time, in minutes) → mean, SD, frequencies, plots, and correlations
3. Fit a null model `m0` predicting `TST`
4. Fit a simple regression model `m1` predicting `TST` by `stress`
5. Fit a multiple regression model `m3` predicting `TST` by `stress` and `insomnia`
6. Compare the three models with the AIC and the likelihood ratio test
7. Print and interpret the coefficients (and their statistical significance) of the selected model
8. Now create two subsets of the `insa` dataset: `insa1` only including observations from the participant `s001` and `insa2` with observations from participant `s002`: how many rows in each dataset?
9. Repeat points 3-7 by using the two subsets: Are results consistent with what you found in the full sample?

# 4. Multilevel data structure

1. Download and read the "*Innovative teaching program*" dataset `studentData.csv` (Moodle/Github, "data" folder)
2. Explore the student-level variables **`studId`** (identification code of each student), **`math_grade`** (student grade in math) and **`anxiety`** (anxiety level). What is the total number of students? How many rows per students? What is the range of **`math_grade`** and **`anxiety`**?
3. How many students per class? How many students per level of the **`tp`** variable?
4. How many classes per level of the **`tp`** variable? To answer that, you can create the **wide-form dataset** by taking only one row per class (e.g., try using the `duplicated()` function preceded by the `!` symbol to remove duplicated values of `classID`)
5. Compute the mean **`math_grade`** and **`anxiety`** value for each class and join them to the wide-form dataset: which is the class with the maximum **`math_grade`**? Which class has the maximum **`anxiety`** level?
6. Fit a simple linear regression model predicting **`math_grade`** by **`anxiety`** both on the long-form and on the wide-form dataset; inspect and interpret the estimated coefficients and their statistical significance.
7. Which model has the highest standard errors? Why?

# 5. Data centering

Consider the long- and wide-form datasets from exercise #4:

1. Compute the **grand-mean-centered** `anxiety` values from the wide-form dataset
2. Fit a simple linear model predicting class-level `math_grade` by grand-mean-centered `anxiety` using the wide-form dataset. Inspect and interpret the estimated coefficients, and compare them with those estimated in the previous exercise
3. Use the `join()` function from the `plyr` package to join the cluster-level mean `anxiety` values to the long-form dataset
4. Compute the **cluster-mean-centered** `anxiety` values by subtracting mean class `anxiety` from student-level `anxiety`
5. Considering class `A`, how many students have an `anxiety` level below the class average? How many have a higher value than the average?
6. Fit a simple linear model predicting student-level `math_grade` by cluster-mean-centered `anxiety` values using the long-form dataset. Inspect and intepret the estimated coefficients, and compare them with those estimated in the previous exercise

# 6. Data centering & level-specific correlations

Do left- and right-side infant pupil sizes correlate more at the within-subject or at the between-subject level?

1. Download and read the "*Infant pupil*" dataset `infantPupil.csv`
2. Subset columns 15, 10, 11, 12, and 13, and rename them as `ID` (subject identification code), `pupil.left` (left-side pupil size in mm), `pupil.left_valid` (validity of left-sized pupil size measurement), `pupil.right` (right-side pupil size in mm), and `pupil.right_valid` (validity of the right-side pupil size measurement)
3. How many valid cases for each eye? (note: 1 = valid, 0 = invalid)
4. Remove all cases with invalid pupil size in either one or the other eye
5. Compute the cluster means and the cluster-mean-centered values for `pupil.left` and `pupil.right`
6. Compute the between-subject and the within-subject correlations between the two variables: Do left- and right-side infant pupil sizes correlate more at the within-subject or at the between-subject level?

# 7. Intraclass correlation coefficient

Using data from exercise #6, compute the intraclass correlation coefficient (ICC) for both pupil size measures.

1. Do they variate more at the within-subject (lv1) or at the between-subject (lv2) level?
2. What is the percentage of within-subject variability over the total variability?
3. Does one eye variate more within-subject than the other?

# 8. Model fit and coefficeint interpretation

1. Download and read the "*Innovative teaching program*" dataset `studentData.csv` (Moodle/Github, "data" folder)
2. Cluster mean center the variable `anxiety` so that we can focus the related slope at the within-individual level
3. Fit a null LMER model `m0` predicting `math_grade`
4. Compute and interpret the ICC
5. Fit a model `m1` with `math_grade` being predicted by `anxiety.cmc`
6. Fit a model `m2` including a random slope for `anxiety.cmc`
7. Fit a model `m3` also including group differences based on `tp` (i.e., innovative teaching program: control vs. intervention) - note: the `tp` variable should be converted as a factor
8. Fit a model `m4` also including the interaction between `anxiety.cmc` and `tp`
9. Inspect and interpret the `summary()` of models `m3` and `m4`
10. What can we say from model `m4`? Does the innovative teaching program improve math achievement? What is the role of anxiety?

# 9. Fixed effect visualization & effect plots

Using models `m1`, `m2`, and `m3` from exercise #8:

1. Visualize and interpret the **forest plot of the fixed effects** by using the `pot_model()` function from the `sjPlot` package
2. Compute the 95% confidence intervals visualized in the plots that you have just generated
3. Visualize and interpret the **fixed effect plots** by adding the argument `type="pred"` within the `plot_model()` function
4. Which are the parameters of model `m1`? Which are those of model `m2` and model `m3`?

# Summary exercises

Exercises #10 and #11 summarize almost all contents that we have seen in Part 1. They are related to the same case study but the former includes some questions similar to those that you might find in the exam, whereas the latter requires you to analyze a multilevel dataset in R.

Note: The reason for which I'm doing most exercises with R is because I find it the most effective way to consolidate the course contents: by concretely analyzing the data, you should better understand the theoretical stuff we see during lectures. However, it is understandable that you also want to get how the exam will be.

# 10. Exam-like exercise: Internet abuse (1/16)

A research is conducted within a project on internet abuse prevention in primary-school children. Data were collected before and after a four-week intervention aiming at (1) reducing children **internet use** (IU), (2) improving the effectiveness of the **parental control** in reducing IU, controlling for children **sex**. Data collection involved 416 children from 14 classes within 7 schools (i.e., two classes per school; one that undertook the intervention and the other that did not).

1. **What is the problem in analyzing such data with linear models (LM)?**

A) There is *no problem* in analyzing such data with LM

B) The local dependencies due to the nested data structure violate the LM assumption of *independence between observations*, possibly biasing the *standard errors*

C) The local dependencies due to the nested data structure violate the LM assumption of *homoscedasticity*, possibly biasing the *t-values*

D) The local dependencies due to the nested data structure violate the LM assumption of *independence between errors and predictors*, possibly biasing the *parameter estimates*

Source: Adapted from Pastore (2021). Analisi dei dati in contesti di comunità

# 10. Exam-like exercise: Internet abuse (2/16)

2. **Why is this a nested data structure?**

A) Because *schools* are nested *within* children

B) Because *children* are nested *between* schools

C) Because the *correlations* among children within the *same school* might be stronger than those between children from *different schools*

D) Because the *correlations* between schools can bias the estimation of the *standard errors*

3. **Which is the dependent variable?**

A) Internet use

B) Children

C) Parental control

D) Intervention

# 10. Exam-like exercise: Internet abuse (3/16)

4. **Which is the cluster variable?**

A) Children

B) Classes

C) Schools

D) Both classes and schools can be considered cluster variables

5. **How many clusters and individual observations?**

A) 7 clusters, 416 individual observations

B) 7 clusters, 14 individual observations

C) 416 clusters, 7 individual observations

D) 416 clusters, 14 individual observations

# 10. Exam-like exercise: Internet abuse (4/16)

6. **Which variables are at the individual-observation level (level 1)?**

A) All variables are at the individual-observation level

B) No variables are at the individual-observation level

C) Internet use, intervention, and sex

D) Internet use, parental control, and sex

7. **Which variables are at the cluster level (level 2)?**

A) All variables are at the cluster level

B) No variables are at the cluster level

C) Intervention

D) Internet use, parental control, and sex

# 10. Exam-like exercise: Internet abuse (5/16)

Here are the first 4 rows of the dataset. `ID` = children identifier, `sex` = children sex (“f” or “m”), `phase` = intervention phase (“pre” or “post”), `IU` = internet use, `CG` = parental control, `school` = school identifier.

```
head(attiva,4)
  ID sex phase IU CG school
1  1   f   pre 22  3      D
2  2   f   pre 11  1      D
3  3   f   pre  9  2      D
4  4   m   pre 13  2      D
```

8. **What kind of dataset are we looking at?**

A) This is the wide-form dataset, with one row per class

B) This is the long-form dataset, with one row per children

C) This is the wide-form dataset, with multiple rows per class

D) This is the long-form dataset, with one row per school

# 10. Exam-like exercise: Internet abuse (6/16)

9. **The between-school correlation between internet use and parental control is -0.1 while the within-school correlation is 0. What does that mean?**

A) A negative (weak) correlation is estimated between school means of internet use and parental control

B) A negative (weak) correlation is estimated between children internet use and parental control within the same school

C) Neither A nor B are true

D) Both A and B are true

10. **How can we compute within-school (level-1) correlations?**

A) We correlate the raw children scores from the long-form dataset

B) We correlate the school mean scores from the wide-form dataset

C) We correlate the cluster-mean-centered scores from the long-form dataset

D) We correlate the grand-mean-centered scores from the wide-form dataset

# 10. Exam-like exercise: Internet abuse (7/16)

```
Loading required package: Matrix
```

11. **The intraclass correlation coefficeint of internet use (IU) is 0.05. What does that mean?**

A) IU mainly variates between schools (95%) than within school (5%)

B) IU mainly variates within school (95%) than between schools (5%)

C) IU equally variates between and within school (50%)

D) None of the previous options is correct

12. **How can we estimate the ICC of a variable?**

A) We divide the estimate of its between-cluster variability by its estimated total variability

B) We subtract the estimate of its between-cluster variability from its estimated total variability

C) We divide the estimate of its within-cluster variability by its estimated total variability

D) We subtract the estimate of its within-cluster variability from its estimated total variability

# 10. Exam-like exercise: Internet abuse (8/16)

13. **Which of the following models is in line with the research goals? (i.e., evaluating intervention reducing internet use (IU) and improving the effectiveness of parental control, accounting for sex)**

A) $IU_{ij} = \beta_1 Intervention + \beta_{2j} Control + \beta_3 Intervention \times Control + \beta_{4j} Sex$

B) $IU_{ij} = \beta_{0j} + \beta_1 Intervention + \beta_{2j} Control + \beta_{3j} Sex + \epsilon_{ij}$

C) $Control_{ij} = \beta_{0j} + \beta_1 Intervention + \beta_{2j} IU + \beta_3 Intervention \times IU + \beta_{4j} Sex + \epsilon_{ij}$

D) $IU_{ij} = \beta_{0j} + \beta_1 Intervention + \beta_{2j} Control + \beta_3 Intervention \times Control + \beta_{4j} Sex + \epsilon_{ij}$

14. **How many fixed coefficients are estimated by that model?**

A) 4: 1 intercept, 3 slopes (intervention, parental control, and sex)

B) 5: 1 intercept, 4 slopes (intervention, parental control, their interaction, and sex)

C) 6: 1 fixed intercept, 1 variance of the random intercept, 4 slopes (intervention, parental control, their interaction, and sex)

D) 7: 1 fixed intercept, 1 variance of the random intercept, 4 slopes (intervention, parental control, their interaction, and sex), 1 residual variance

# 10. Exam-like exercise: Internet abuse (9/16)

Here are the results of a first model that only tests whether the intervention was able to reduce internet use, controlling for sex.

| Predictors | b (SE) | CI | p |
|---|---|---|---|
| (Intercept) | 9.45 (0.59) | 8.28 – 10.62 | **<0.001** |
| phase [post] | -0.98 (0.41) | -1.78 – -0.18 | **0.016** |
| CG | 1.96 (0.30) | 1.37 – 2.55 | **<0.001** |
| sex [f] | 0.20 (0.44) | -0.67 – 1.06 | 0.656 |
| Random Effects | | | |
| $\sigma^2$ | 16.92 | | |
| $\tau_{00}$ school | 0.49 | | |
| $\tau_{11}$ school.CG | 0.29 | | |
| $\rho_{01}$ school | 0.07 | | |
| N school | 7 | | |
| Observations | 412 | | |

Note: `sjPlot` calls random effect variances $\tau$ rather than $\tau^2$

# 10. Exam-like exercise: Internet abuse (10/16)

Considering the table shown in the previous slide:

15. **Was the intervention (variable `phase`) effective in reducing internet use?**

A) No, because the coefficient $\beta$ = 1.96 is positive

B) Yes, because the coefficient $\beta$ = 1.96 is not lower than $t$ = 1.96

C) No, because the coefficient $\beta$ = -0.98 is lower than $t$ = -1.96

D) Yes, because the coefficient $\beta$ = -0.98 has a $p$-value lower than 0.05

16. **What random effects are reported in the table?**

A) Residual variance, random intercept (RI), and random slope (RS)

B) Residual variance, RI, RS, and correlation between RI and RS

C) Residual variance, RI, RS, and ICC

D) Residual variance, RI, RS, correlation between RI and RS, and ICC

# 10. Exam-like exercise: Internet abuse (11/16)

Considering the table shown in the previous slide:

17. **How can we interpret the coefficient estimated for `phase`?**

A) In the post-intervention group of children, the model predicts an IU value 0.98 points lower than in the pre-intervention group

B) For a one-unit increase in `phase`, the model predicts a decrease in IU by 0.98

C) In the post-intervention group of children, the model predicts an IU value 0.98 points higher than in the pre-intervention group

D) For a one-unit increase in `phase`, the model predicts an increase in IU by 0.98

18. **If we trust the p-values, which fixed effects are significant?**

A) Intercept, phase, and sex

B) Intercept, phase, and CG

C) Intercept, phase, CG, and random intercept

D) Intercept, phase, sex, and random intercept

# 10. Exam-like exercise: Internet abuse (12/16)

Here are the results of a second model that additionally tests whether the intervention was able to improve the effectiveness of the parental control in reducing internet use, controlling for sex.

| Predictors | b (SE) | CI | p |
|---|---|---|---|
| (Intercept) | 9.00 (0.68) | 7.66 – 10.34 | **<0.001** |
| phase [post] | 0.19 (0.89) | -1.55 – 1.94 | 0.830 |
| CG | 2.23 (0.40) | 1.45 – 3.02 | **<0.001** |
| sex [f] | 0.14 (0.44) | -0.71 – 1.00 | 0.741 |
| phase [post] × CG | -0.64 (0.43) | -1.48 – 0.20 | 0.137 |
| Random Effects | | | |
| $\sigma^2$ | 16.96 | | |
| $\tau_{00}$ school | 0.00 | | |
| $\tau_{11}$ school.CG | 0.42 | | |
| $\rho_{01}$ school | | | |
| N school | 7 | | |
| Observations | 412 | | |

# 10. Exam-like exercise: Internet abuse (13/16)

Considering the table shown in the previous slide:

19. **What coefficient has been added in this scond model compared to the first one?**

A) There is a new coefficient for the random slope for parental control

B) There is a new coefficient for the main effect of the intervention

C) There is a new coefficient for the interaction between intervention and parental control

D) The two models are equivalent

20. **Considering the significance of the coefficients shared by both models, what has changed?**

A) The coefficients estimated for phase and CG are no longer significant

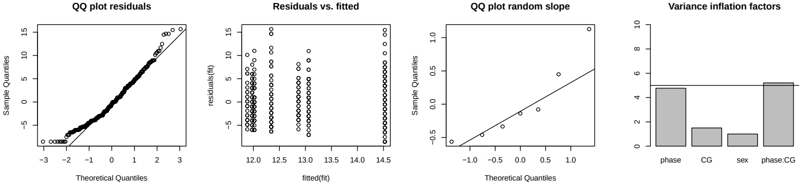
B) The coefficient estimated for phase is no longer significant

C) The coefficient estimated for CG is no longer significant

D) Nothing has changed in the significance of the shared coefficients

# 10. Exam-like exercise: Internet abuse (14/16)

Here are some plots showing the diagnostics of the second model.



21. **Which LMER assumptions are evaluated by these plots?**

A) Normality and linearity of residuals

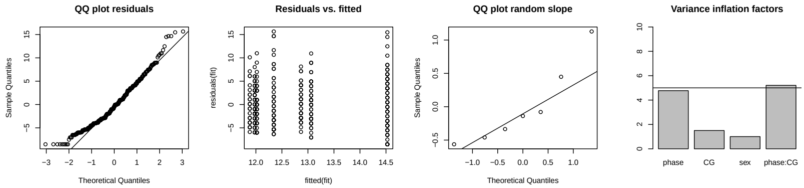
B) Independence and homoscedasticity of residuals

C) Normality of random slope and absence of multicollinearity

D) All other options are true

# 10. Exam-like exercise: Internet abuse (15/16)

Here are some plots showing the diagnostics of the second model.



22. **Which LMER assumptions are violated based on these plots?**

A) Normality is violated for both residuals (slightly) and random slopes (more evident)

B) Homoscedasticity is violated for both residuals (more evident) and random slopes (slightly)

C) Normality is violated for both residuals (slightly) and random intercepts (more evident)

D) Normality of residuals (more evident) and absence of influential cases (slightly)

# 10. Exam-like exercise: Internet abuse (16/16)

Now we repeat the analysis described in the previous slides by using R. First, download and read the `attiva.RData` dataset (note `rda` files can be read similar to `RData` files).

The likelihood ratio test and the Akaike Information Criterion (AIC) were used to compare the two models. The former resulted in a $\chi^2$ value of 1.52 that with 1 degrees of freedom is equivalent to a *p*-value of 0.22. The latter resulted in an AIC value of 2361.85 for the first model and 2362.3 for the second model.

23. **What conclusion can be drawn from the likelihood ratio test results?**

A) The first (simpler) model is better than the second (more complex) model

B) The second (more complex) model is better than the first (simpler) model

C) The two models are not different

D) None of the preceding options are true

24. **What conclusion can be drawn from the AIC results?**

A) The first (simpler) model is better than the second (more complex) model

B) The second (more complex) model is better than the first (simpler) model

C) The two models are not different

D) None of the preceding options are true

# 11. R-based summary exercise

Now we repeat the same steps described in the previous slides by using R.

1. Download and read the `attiva.RData` dataset from Moodle or Github: `ID` = children identifier, `sex` = children sex ("f" or "m"), `phase` = intervention phase ("pre" or "post"), `IU` = internet use, `CG` = parental control, `school` = school identifier.
2. Explore the data (mean, SD, frequencies, and plots)
3. Compute the level-specific correlations between
4. Create the variable `class` (identifying the class) by crossing `phase` and `school`. Then, compute the ICC of IU based on `school` and that based on `class`. What is the best cluster variable to be used?
5. Fit a random-intercept model `m1` including `phase`, `CG`, and `sex` to be compared with a second model `m2` that also include the random slope for `CG` using the likelihood ratio test and the AIC. Which is the best model?
6. Add the interaction between `phase` and `CG` to the model selected above and compare these two models with the likelihood ratio test and the AIC. Which is the best model?
7. Evaluate the diagnostics of the selected model
8. Print, visualize and interpret all the effects (fixed and random) estimated by the selected model

# 12. Variance, covariance, & correlation

1. Simulate 100 values of a normally distributed variable `x1` with mean 10 and standard deviation 2 (note: use the `rnorm(n,mean,sd)` function where `n` is the sample size, `mean` is the simulated variable mean, and `sd` is the simulated variable standard deviation), then plot `x1`
2. Simulate 100 values of a second normally distributed variable `x2` with mean 10 and standard deviation 10, then plot `x2`
3. Try answering before running the code: Which variable has the highest **variance**?
4. Simulate a third variable `x3` by adding a small random quantity to `x1`: `x3 <- x1 + rnorm(n = 100, mean = 0, sd = 1)`, then plot `x3`
5. Simulate a fourth variable `x4` by adding a large random quantity to `x1`: `x4 <- x1 + rnorm(n = 100, mean = 0, sd = 20)`, then plot `x4`
6. Try answering before running the code: Which variable has the highest variance?
7. Try answering before running the code (but you can try plotting the variables): Which variable has the highest **covariance** with `x1`? Which has the lowest covariance?
8. Try answering before running the code: Which variable has the highest **correlation** with `x1`? Which has the lowest correlation?

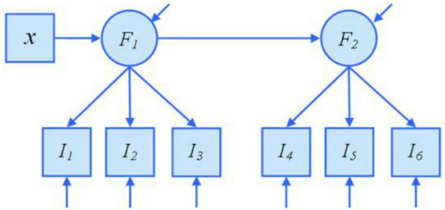
# 13. Covariance & correlation matrix

1. Download and read the *Pregnancy during pandemics* data, then select the columns 2, 5, and 14, and rename them as “age”, “depr”, and “threat” (see `2-multilevel.pdf` slide #10)
2. Use the function `na.omit()` to remove all cases with missing values in one or more variables. How may observations have been removed?
3. Describe and plot the included variables
4. Compute the **covariance matrix** $S$ of the variable and use the logic operators `==` (is equal) and `!=` (is different) operators to check the two properties of the covariance matrix S (symmetry and main diagonal)
5. Standardize all variables ($z_{x_i} = (x_i - \overline{x})/s_x$ ), plot them, and compute their mean and SD. What do you observe?
6. Re-compute the covariance matrix on standardized variables. What do you observe?
7. Compute the **correlation matrix** of the original variables and compare it with the covariance matrix computed from standardized variables. What do you observe?

# 14. Reading a SEM plot

In the figure below:

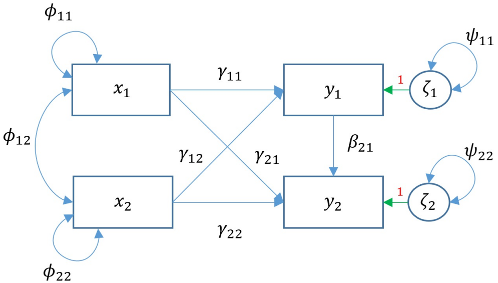
1. How many **latent & observed** variables? Which ones?
2. How many **endogenous & exogenous** variables? Which ones?
3. How many **error terms**? Which ones? (note: this plot represents errors/residuals by using the third graphical notation shown in `3-multivariate.pdf` slide #21)
4. How many **path coefficients** (i.e., single-headed arrows)? Which ones?
5. Which path coefficients are included in the **measurement model**? Which ones in the **structural model**?
6. How many estimated parameters, in total?



# 15. Reading a path diagram (1/2)

In the figure below (which indexes parameters using different letters than those we saw in class, but don't care about that):

1. How many **latent & observed** variables? Which ones?
2. How many **endogenous & exogenous** variables? Which ones?
3. How many **error terms**? Which ones? (note: this plot represents errors/residuals by using the first graphical notation shown in `3-multivariate.pdf` slide #21)
4. How many **path coefficients** (i.e., single-headed arrows)? Which ones?
5. How many **covariances**?
6. How many estimated parameters, in total?



# 15. Reading a path diagram (2/2)

Which system of equations would you use to represent the model shown in the previous slide? (note 1: for better consistency, I'm using the same letters used in the figure, but don't care about it; note 2: variances and covariances are not reported in the equations)

A)

$$\begin{cases} y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \zeta_2 \\ y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \zeta_1 \end{cases}$$

B)

$$\begin{cases} y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \zeta_2 \\ y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \zeta_1 \\ x_2 = \beta_1 x_1 + \gamma_{12}y_1 + \gamma_{22}y_2 + \epsilon_{x1} \end{cases}$$

C)

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \beta_{21}y_1 + \zeta_2$$

D)

$$\begin{cases} y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \beta_{21}y_1 + \zeta_2 \\ y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \zeta_1 \end{cases}$$