Recap session

Multivariate Statistics

2024

General questions

Terminology

Nominal and ordinal can also be summarized as categorical variables, right?

Yes, nominal and ordinal variables are summarized as categorical variables.

Interval scaled variables are also called numeric variables or metric variables?

Interval-scaled variables are often referred to as numeric or metric variables.

Then we also differ between continuous and discrete variables of which discrete variables mean interval scaled because there are clear spaces between the values. Continuous variables mean that the data is continuous or what we would call a ratio scale. Is this correct?

Discrete variables do not necessarily mean interval-scaled; they refer to variables with distinct, separate values. Continuous variables do not always mean ratio scale but rather indicate unbroken data within a range. Ratio scale is a type of continuous variable.

Interaction and control variables

What [are] interactions and control variables?

Interactions: Help capture more complex relationships by modeling how the effect of one independent variable on the dependent variable changes depending on the level of another variable.

Control variables: Help isolate the relationship between the main independent variables and the dependent variable by accounting for confounding factors that might influence the outcome.

Linear regression

Multivariate vs multiple regression

We looked at multivariate models which you defined as the influence of third variables on the relation of two variables. However, in the subchapter relationship between r and R-square we are talking about multiple regression […] Could explain the difference between multivariate and multiple regression once more?

In the subchapter discussing the relationship between r and R^2, the focus is on multiple regression as it involves one dependent variable (Y) and multiple independent variables (X1, X2, etc).

The website refers to "multivariate models" (strict definition: models with multiple dependent variables) in a way that aligns with "multiple regression", emphasizing the inclusion of third (or more) variables to:

- Account for their influence on the primary relationship of interest.
- Estimate "net" effects, correct for spurious relationships, and detect interaction and indirect effects.

The website seems to use "multivariate" colloquially to refer to multiple regression models that control for additional variables, which is common in applied contexts.

Leverage

Could you explain the concept of "leverage" as a factor for the impact of outliers again?

Leverage is a concept in linear regression that measures how far an individual data point's value for the independent variable (X) is from the mean of X, relative to all other observations. It quantifies the potential influence a data point has on the regression line because of its position in the X-space.

- Leverage alone does not determine whether a data point is an outlier. A data point with high leverage has the potential to **influence** the regression line, but its actual impact depends on whether its Y-value aligns with the regression line.
- If a high-leverage point also has a large **residual** (the difference between the observed and predicted Y-value), it becomes a high-influence outlier. Points with high leverage but small residuals may not substantially affect the regression line.

In practice, leverage values are compared to a threshold to identify potentially problematic points:

$hi > 2(k+1) / n$

k: Number of predictors in the model
n: Total number of observations

Limitation(s) of R2

[What are] limitations of R2?

- measures the proportion of variance explained by the linear model. It does not capture the quality of fit in non-linear relationships.
- A high R^2 does not imply a causal relationship between independent and dependent variables. It only indicates the strength of the association.
- In models with a large number of predictors, R^2 tends to increase regardless of whether those predictors are meaningful.
- R^2 cannot account for the appropriateness of the predictors, the presence of omitted variable bias, or the influence of irrelevant variables.
- Outliers can inflate or deflate R^2, giving a misleading picture of the model's explanatory power.
- R^2 measures the proportion of explained variance but does not provide information on prediction error or accuracy.
- R^2 assumes that the variance of residuals (errors) is constant. If this assumption is violated (heteroscedasticity), R^2 might not accurately represent model performance.
- For some types of regression models, such as logistic regression, R^2 is not meaningful or applicable in its standard form. Alternatives like McFadden's R^2 or other pseudo-R^2 metrics are used instead.
- Overfitted models may show an inflated R^2 on training data, but this "shrinks" (as a decline in model performance on unseen data) when applied to new data because irrelevant predictors or noise contribute less to predictive power.

Standard error

How well do we need to understand standard estimation error?

Understanding the standard error of estimation (or standard error of the regression) is crucial for evaluating the precision and reliability of a regression model.

The standard error measures the average distance that the observed values fall from the regression line. A smaller standard error indicates that the data points are closer to the regression line, suggesting a more precise model.

The standard error is essential for constructing confidence intervals around regression coefficients and predictions.

In regression analysis, the standard error is used in t-tests to determine if a predictor's coefficient is significantly different from zero.

Normality of residuals and homoscedacity

"Heteroscedasticity occurs when the variability of the residuals (errors) depends on the estimated Y values (predicted values of Y)." I confuse normality of residuals and homoscedacity.

Heteroscedasticity is about the spread of residuals across predicted values. If the spread increases (or decreases), the residuals are heteroscedastic

= > Can lead to biased estimates of standard errors, making hypothesis tests unreliable.
=> Use scatter plot of residuals vs. predicted values.

Normality is about the shape of the residuals' distribution. It does not depend on predicted values but rather looks at how residuals are distributed overall

=> Can affect the validity of p-values and confidence intervals but does not typically impact predictions.
=> Use histogram or Q-Q plot of residuals.

Interactions

Please explain interactions of interval variables and dummy variables.

The interaction captures whether and how the slope of the interval variable (X) differs across categories of the dummy variable (D).

It allows the effect of the interval variable to vary depending on the category of the dummy variable.

Graphical representation:

- No Interaction: The relationship between X and Y (slope) would be the same for categories of D (parallel lines)
- With Interaction: The slope of X (interval variable) differs between the categories of D (lines intersect or diverge)

Logistic regression

Difference between p, odds, odds ratio, log odds (logit)

Until now I don't really get the difference between p, odds, odds ratio, log odds (logit), and probabilities and what they say about the probable relation between two variables.

Probabilities (p) represent the likelihood of an event occurring, ranging from 0 to 1.

Odds compare the probability of an event occurring to it not occurring: $p/(1-p)$. Odds ratio compares the odds of an event between two groups, showing relative likelihood.

Log odds (logit) is the logarithm of odds, used in logistic regression to linearize probabilities for modeling.

Marginal predictions

Could you explain again the concept of marginal predictions in logistic regression and clarify at which step of the analysis they are applied?

Marginal predictions in logistic regression represent the predicted probabilities (P(Y=1)) for different values of a specific independent variable, while holding all other independent variables constant (typically at their mean, median, or another fixed value).
Marginal predictions are derived at this final step, where probabilities are available, and allow for variation in one variable while holding the others constant.

- For **continuous** variables: Marginal predictions show how P(Y=1) changes as the continuous variable (X) changes incrementally, controlling for other variables.
- For **dichotomous** variables: Marginal predictions calculate the difference in adjusted probabilities for the two groups of a binary variable (e.g., men vs. women).
- For **discrete** variables: Marginal predictions show the effect of discrete changes in a categorical variable on P(Y=1), such as a shift from "high school" to "some college" to "college degree."

Non-linear model

Logistic regressions are non-linear probability models, right?

Logistic regression models the relationship between a set of independent variables (X) and a binary dependent variable (Y).The key difference from linear regression is that the predicted outcome in logistic regression is a probability (P(Y=1)), which must lie between 0 and 1.

The relationship between the independent variables and the predicted probability is non-linear because the model applies a logistic (sigmoid) transformation to ensure the probabilities fall within this range.

The output of logistic regression is the predicted probability (P(Y=1)) of the binary outcome.

ANOVA and repeated measures (RM) ANOVA

University of Zurich[UZH]

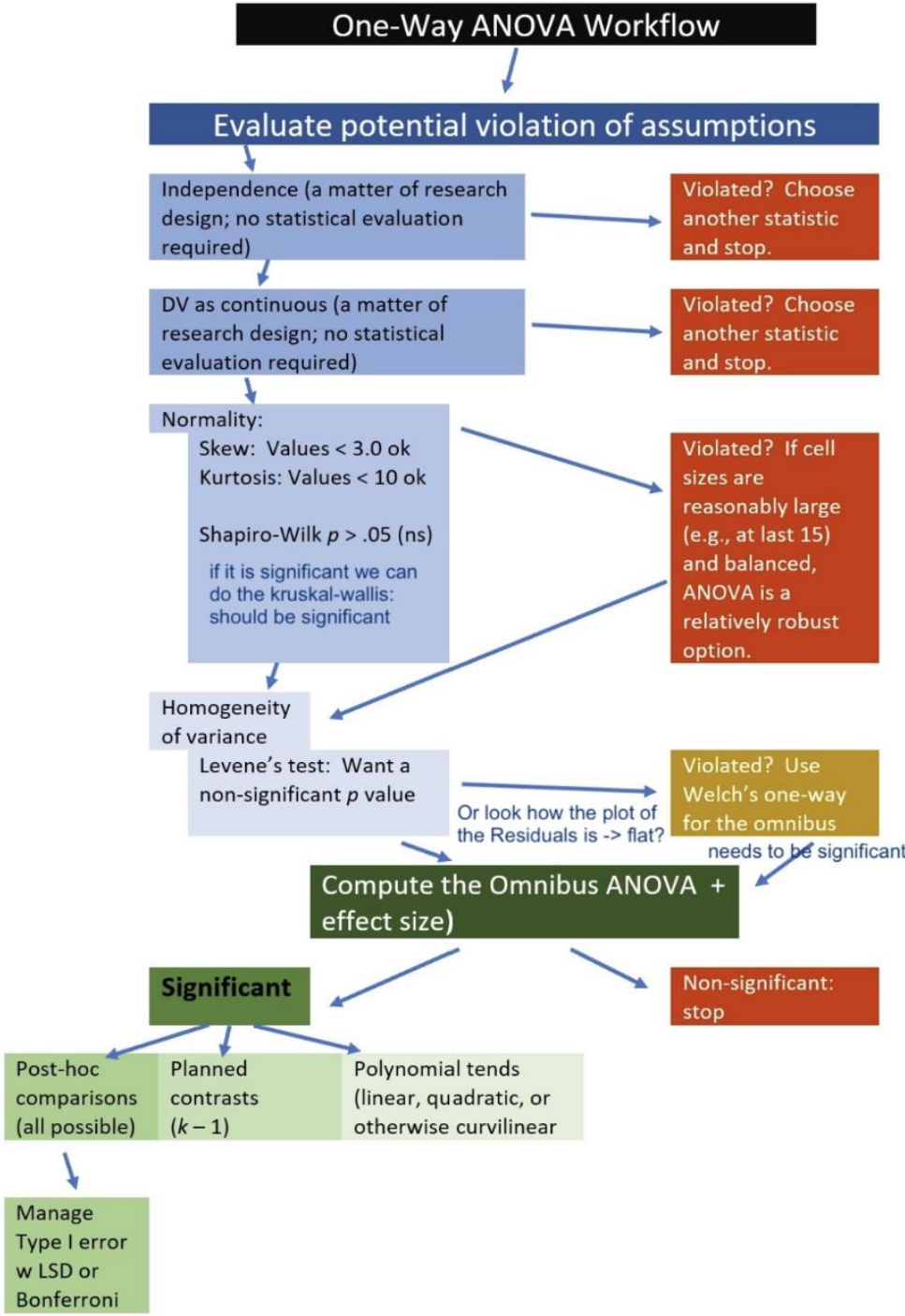
Assumption testing for ANOVA

I'm struggling with the assumption testing for the ANOVAs. First, I see that there are many different tests (Mauchly, Kruskal, visual testing, Welch, etc.). Some need to be significant, some not. And sometimes, even if they don't fit, we continue with the analysis. Is there a logical way, or order to process these tests?

The process outlined in the image is logical and reflects the correct order of assumption testing for ANOVAs:

- independence and continuity (based on design, not tests)
- normality (e.g., Shapiro-Wilk, skew, kurtosis)
- homogeneity of variance (e.g., Levene's test)

If assumptions are violated, alternative approaches like Kruskal-Wallis or Welch's ANOVA are suggested, which aligns with best practices.



Difference between eta-squared and partial eta-squared

Eta-squared: Proportion of variance (TSS) explained by the factor (SSeffect), including overlap with other factors/variables (for simple one-way ANOVA designs with only one factor).
Partial eta-squared: Proportion of variance (TSS) explained by the factor (SSeffect), after controlling for other factors (for complex designs with multiple factors or interactions).
Both range between 0 and 1, with values closer to 1 indicating a stronger effect or better model fit.
Does this sound correct?

Eta-squared is a measure of the proportion of the total variance in the dependent variable (TSS) that is explained by a particular factor (SSeffect). It includes variance overlap with other factors or variables in the model. In simple one-way ANOVA, where there is only one factor, eta-squared provides a straightforward measure of effect size.

Partial eta-squared is a measure of the proportion of variance in the dependent variable that is uniquely attributable to a specific factor (SSeffect) after controlling for other factors or interactions. It isolates the variance explained by a single factor by excluding the shared variance with other factors. It is used in more complex designs, such as factorial ANOVA or repeated measures designs, where multiple factors or interactions are involved.

Eta-squared equivalent to R^2 in one-way ANOVA

Is it correct that eta squared (η^2) is equivalent to R^2 in the context of one-way ANOVAs?

Eta Squared (η2) in one-way ANOVA measures the proportion of total variance (SStotal) in the dependent variable that is explained by the independent variable (the factor).

In linear regression, R2 measures the proportion of total variance in the dependent variable that is explained by the predictors.

In one-way ANOVA, the independent variable is categorical (representing group membership), and the dependent variable is continuous.

The ANOVA model can be expressed equivalently as a regression model, where the categorical variable is dummy-coded. The proportion of variance explained (η2) in ANOVA matches R2 from the regression.

RSS

In Anova, RSS tells you how much variability exists within each group. It's a measure of the error variance or the variance not explained by the group differences. How does this relate to alpha – cumulated errors if at all?

RSS represents the variability within groups that cannot be explained by the differences between the groups. RSS directly impacts the F-statistic in ANOVA, affecting whether group differences are statistically significant given the alpha level.

Alpha controls the threshold for Type I errors (when a true null hypothesis is incorrectly rejected), and adjustments to alpha (e.g., in multiple comparisons) are influenced indirectly by RSS because RSS affects the precision of post hoc tests.

While not directly tied to alpha or cumulated errors, RSS plays a critical role in determining the likelihood of detecting significant differences under the chosen significance level.

ANCOVA

1) We can use ANOVA to test if mean test scores differ between students from 3 schools. But if I want to control for gender, I cannot use ANCOVA due to type of control variable - is this correct? What shall I use?

ANCOVA is not suitable when a categorical control variable (like gender) interacts with the main grouping variable.

Use a Two-Way ANOVA or a regression model with interaction to appropriately analyze the data and control for gender while allowing for interactions between school and gender.

2) In ANCOVA, How can we detect outliers? Through plots only?

Plots provide an intuitive way to identify data points that deviate significantly from the expected relationships.

Examples:

- Residual plot: Plot residuals against predicted values or covariates
- Scatter plot with covariate: Plot the dependent variable against the covariate, colored or separated by groups
- Cook's distance: Measures the influence of each observation on the regression coefficients
- Leverage statistics: Identifies observations with unusual predictor values

Moderation and mediation analyses

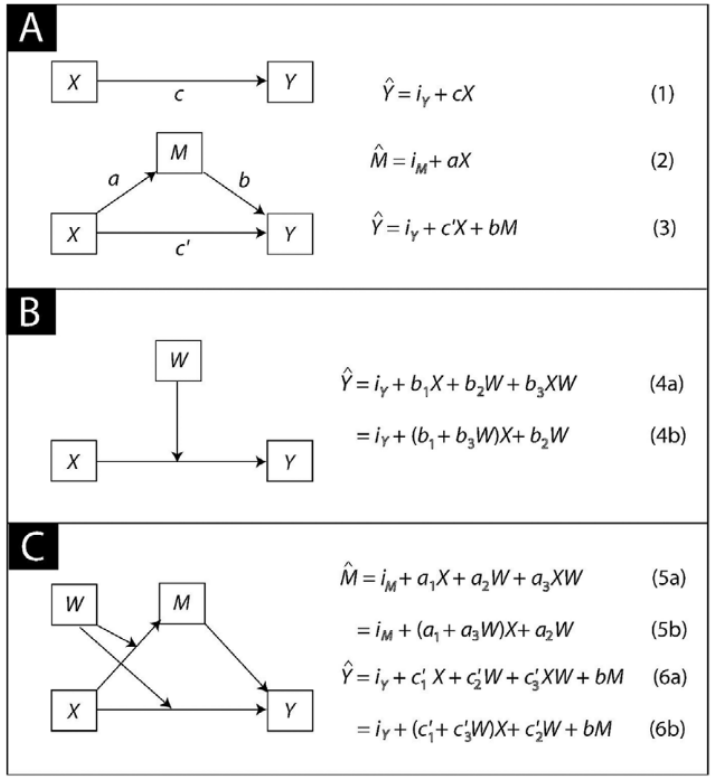


Figure 1. Simple Mediation (A), Simple Moderation (B), and a Conditional Process Model (C)

Igartua, J. J., & Hayes, A. F. (2021). Mediation, moderation, and conditional process analysis: Concepts, computations, and some common confusions. *The Spanish Journal of Psychology*, 24, e49.

Simple and conditional effects

Is it possible to explain the simple main effect (conditional effect) one more time? Also referring to slides, section 2.4 "Correct Interpretations".

In moderation analysis, a simple main effect (or conditional effect) refers to the relationship between the independent variable (X) and the dependent variable (Y) at a specific level of the moderator (M).
Essentially, it shows how X affects Y when M is held constant at a particular value.

2.4 Correct interpretations

- b_1 is the conditional effect of X on Y when Z = 0.
- b_1 is the estimated difference in Y between two cases in the data that differ by one unit in X but have a value of 0 for Z.
- b_2 is the conditional effect of Z on Y when X = 0.
- When X = 0, the conditional effect of Z reduces to b_2.
- When Z = 0, X's effect equals b_1; when Z = 1, X's effect equals $b_1 + b_3$; when Z = 2, X's effect is $b_1 + 2b_3$, and so forth.

EFA and CFA

EFA: choosing the "correct" number of factors

How to choose how many factors to proceed with when screeplot and eigenvalues show a different number? Is one method "stricter" than the other?

When the scree plot and eigenvalues suggest different numbers of factors, the choice depends on context:

- Eigenvalues > 1 (Kaiser criterion) can overestimate factors, as it's less strict.
- The scree plot, focusing on the "elbow" where the slope flattens, is stricter and emphasizes parsimony.

Generally, the scree plot is preferred when the two methods conflict, but theoretical justification and interpretability should guide the final decision.

CFA: should we do CFA after EFA?

In the lesson we learned that it doesn't make sense to do a CFA after a EFA but this is written on the website: "Recommendation: It is best to have an EFA before the CFA - either with the same or with a different sample." Was it perhaps a special case in the exercise lesson and normally CFA and EFA should be done in combination?

EFA (Exploratory Factor Analysis) is typically used when you are in the early stages of your research and have no strong theoretical basis or prior knowledge about the factor structure of your constructs.

CFA (Confirmatory Factor Analysis), on the other hand, is used when you already have a hypothesized model based on theory or prior research.

A) **EFA before CFA is recommended when:** Performing an EFA ensures that your factor structure is empirically grounded, reducing the likelihood of specifying an incorrect model for CFA.

B) **EFA before CFA is <u>not</u> recommended when:**
   - Conducting both EFA and CFA on the exact same dataset can lead to overfitting and artificially inflated results.
   - A theoretical model is already well-established, you might skip EFA and directly conduct CFA.

EFA/CFA: item and indicator are interchangeable?

Can we use the terms item and indicator interchangeably, or what exactly is the difference?

The terms "item" and "indicator" are often used interchangeably. There are subtle differences in their meanings depending on the context:

- An **item** generally refers to a single question, statement, or prompt on a survey, questionnaire, or test.
- An **indicator** is a broader term that refers to any observable variable used to reflect or "indicate" an underlying latent construct.

Aspect	Item	Indicator
Focus	Specific question or statement.	Observable variable reflecting a construct.
Granularity	Refers to survey or test questions.	Can include items and other measures.
Use in Models	Used to collect data for indicators.	Used in models like EFA/CFA to define latent variables.
Context	Common in surveys and tests.	Common in factor analysis and SEM.

CFA: indicator correlations

With respect to the mathematical requirements of CFA, the mock exam says that indicators should not correlate with one another but on the website (chapter 10.2) it says that they should correlate with one another. Which one is correct?

Website content: states that items that should theoretically load on a factor should correlate empirically. This implies that indicators (items) associated with the same factor are expected to correlate because they reflect the shared variance due to the underlying latent construct.

Mock-exam content: asks "In confirmatory factor analysis (CFA), which of the following statement is NOT a mathematical requirement?" and the correct answer is "Items that should theoretically load on a factor should not correlate empirically." This is because if indicators did not correlate, it would suggest they are unrelated and not measuring the same construct.

CFA: parameters and identification

1) Could you explain again what a parameter is and how I evaluate if it is a fixed, constrained or free one?

A parameter in CFA is a value or coefficient in the model that represents some aspect of the relationships between variables, including: factor loadings (how strongly each observed variable (indicator) is associated with the latent factor); covariances (relationships between latent factors or between observed variables); residuals (variances of the measurement error in observed variables).

- A parameter is **fixed** if its value is set by the researcher and not estimated from the data (e.g., one factor loading is often fixed to 1 to set the scale of the latent factor)
- A parameter is **constrained** if it is restricted to have the same value as another parameter or is restricted within a certain range (e.g., constraining two factor loadings to be equal to test whether indicators have equivalent contributions to the factor)
- A parameter is **free** if its value is estimated from the data without constraints

2) If I have an underidentified model, how exactly can I gain more degrees of freedom to make it an identified one?

An underidentified model has fewer degrees of freedom (df) than needed to estimate all free parameters. Solutions are:

- Reducing the number of free parameters: fixe or constrain some parameters
- Add more observed indicators per latent factor
- Simplify the model
- Use appropriate identification rules: each latent factor has => 3 observed indicators; identification method

Structural equation modelling (SEM)

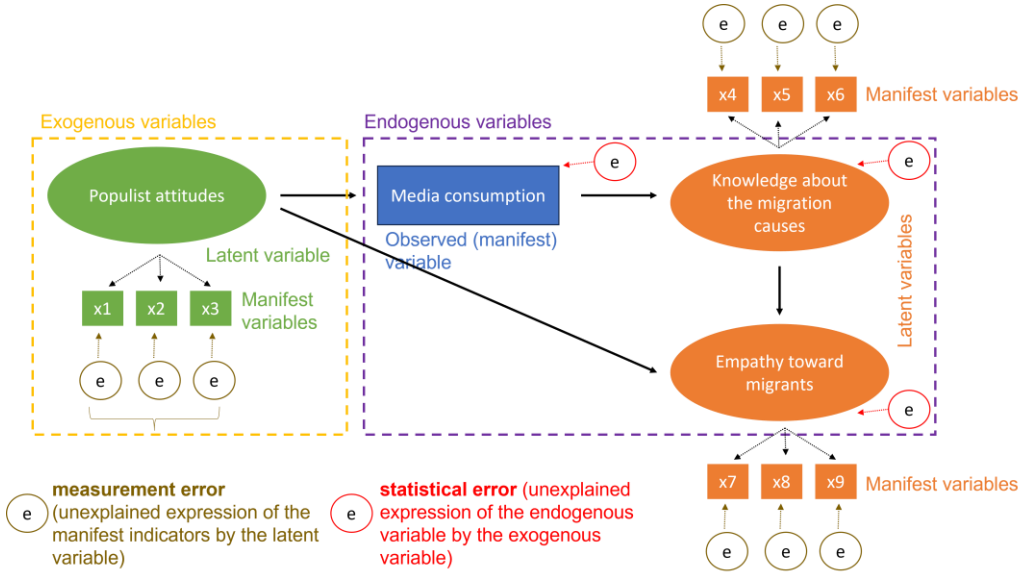
New 10/12

Statistical errors and measurement errors (I)

Could you explain the difference between statistical errors and measurement errors?

Statistical errors are deviations between the predicted values from a model and the actual observed values. These arise from the inability of the model to perfectly predict or explain the observed outcomes.

Measurement errors refer to inaccuracies in the observed values of variables due to imperfections in the measurement process. These are discrepancies between the true (latent) variable and the observed (measured) variable.



New 12/12

Statistical errors and measurement errors (II)

Also you wrote for SEM that there is a measurement model and a structural model would that refer to the theoretical one and the actually found empirical one through data, like in CFA? I understand recursive as an iterative process, but did not get why there would be repercussions?

The measurement model and structural model are components of your theoretical model but are evaluated empirically (like in CFA).

Optional:

Recursive models are simpler and avoid the complications of feedback or simultaneous causality.

Non-recursive models require attention to potential repercussions because of the complexity of feedback loops, which can complicate interpretation and estimation.

Multilevel modelling (MLM)

New 10/12

RM-ANOVA versus MLM

To me it seems similar to Repeated Measures ANOVA. For example, the case we saw that compared the individuals and the results between different countries. When do we use Repeated Measures ANOVA and when Multilevel modelling?

Repeated Measures ANOVA (RM-ANOVA) and Multilevel Modeling (MLM) can be used for analyzing data with repeated measures, but they differ in flexibility and assumptions:

Some key differences:

Aspect	**Repeated Measures ANOVA**	**Multilevel Modeling**
Flexibility	Limited; assumes sphericity and complete data	High; handles missing data and complex structures
Data Structure	Balanced and simple	Nested or unbalanced data
Random Effects	Limited to within-subject effects	Allows random intercepts and slopes
Covariates	Limited options	Can include level-specific covariates

New 13/12

ICC

When do we use the ICC ? With the null model? Or with a random intercept model?

The ICC is a key statistic in multilevel modeling that quantifies the proportion of variance in the outcome variable attributable to the grouping structure (e.g., schools, countries, etc).

Compute the ICC with the null model to understand the baseline proportion of variance due to clustering.

If you compute the ICC with a random intercept model, it reflects residual clustering effects after accounting for predictors, which is less common for reporting the ICC but can be insightful for interpreting model refinement.

Supplementary material from 2023

Linear regression

Dummy variables

What is the reference category - 0 or 1?

The reference category is typically 0 (by default in R).

How many characteristics do categorical variables need to have to work as dummy variables?

For a categorical variable with k distinct categories, you generally need k−1 dummy variables to represent all the categories effectively.

Reminder:

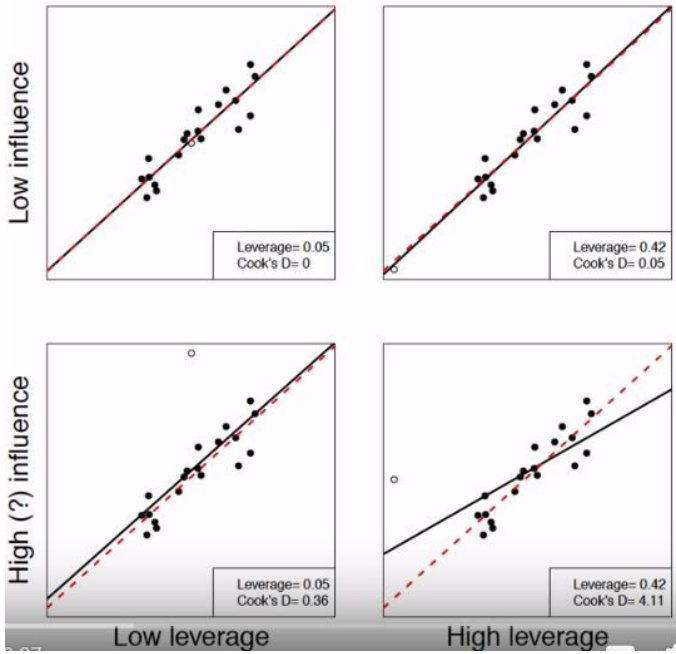
Dummy variables enable the modelling of the impact of different categories of a variable on the outcome variable by estimating separate coefficients for each category relative to a chosen reference category.

Outliers and leverage

Reminder:

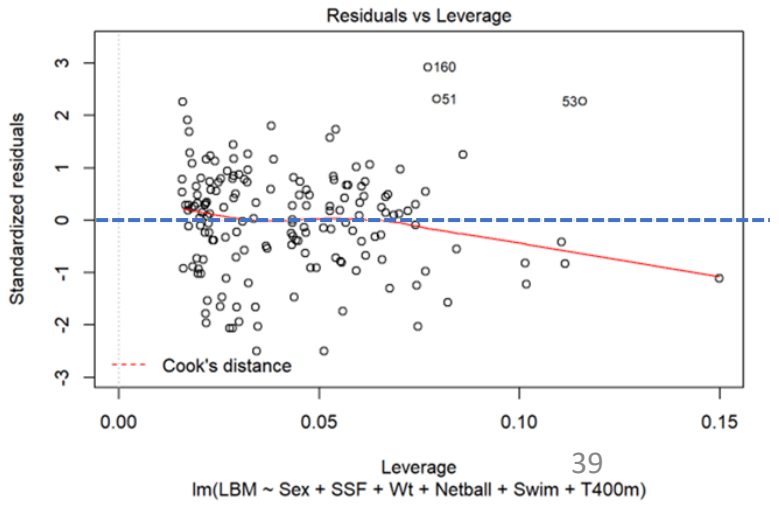
Outliers are data points that are **unusually distant from other observations in the dependent variable >>** contribute to larger errors between observed and predicted values **>>** can skew the results and **affect the overall goodness-of-fit** of the regression model.

High leverage points are observations that have **extreme predictor values >>** might not necessarily have a large residual (difference between observed and predicted values) **>>** can greatly impact the regression line's slope and position, thus **affecting the model fit and predictive capabilities**.



And what do we want to see on a residuals vs leverage plot?

In a **residuals vs leverage plot**, you want to see most of the points clustered around zero residual lines (horizontal lines at y=0), indicating that the residuals are not systematically high or low across different levels of leverage. The plot helps in identifying influential points that might significantly affect the regression analysis.



Assumptions / diagnostics

What does mean independence refer to (as an assumption)?

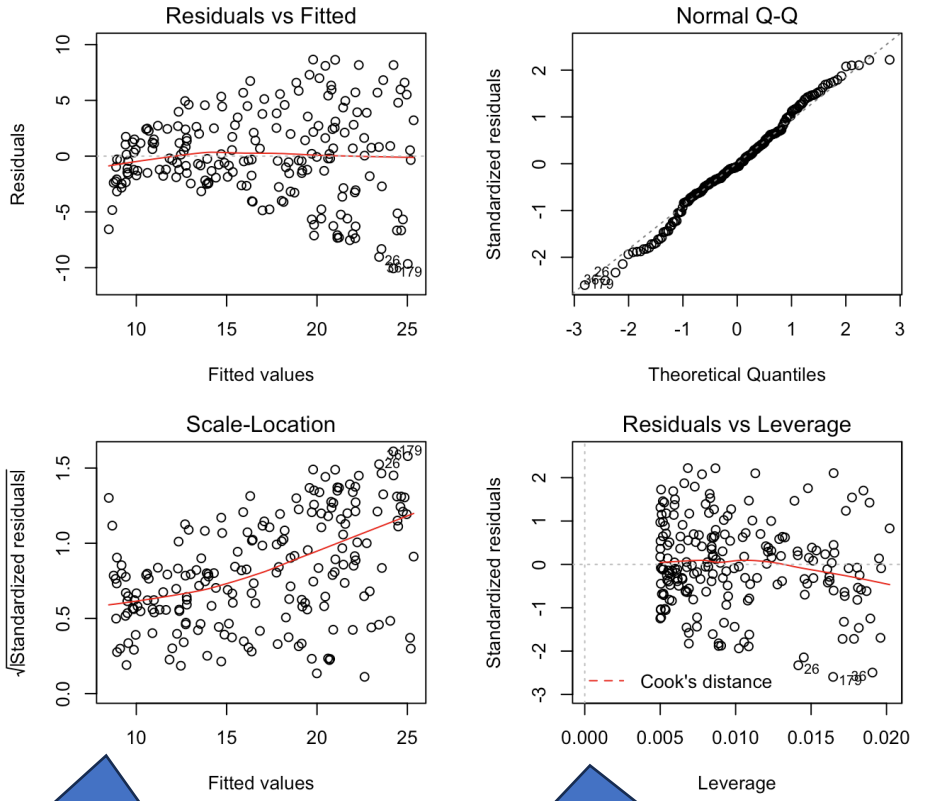
In linear regression, the assumption of independence refers to the **independence of the errors or residuals**.

This assumption states that the errors or residuals (the differences between observed and predicted values) are independent of each other.

Assuming independence of errors is crucial because **violating this assumption can lead to biased estimates, incorrect standard errors, and unreliable statistical inferences**.

check the linear relationship assumptions

examine whether the residuals are normally distributed



check the homogeneity of variance of the residuals (homoscedasticity)

identify influential cases

Standardized coefficients

Can we say that standardized b shows how Y varies when X increases by 1? Or is it only for comparing the effects of independent variables after standardization?

Both unstandardized (b) and standardized (Beta) coefficients provide valuable information in interpreting the effect of independent variables on the dependent variable:

- The **standardized coefficient (Beta)** helps in comparing the relative impact of predictors.
- The **unstandardized coefficient (b)** directly relates to the change in the dependent variable for a one-unit change in the independent variable in its original units.

Example: Standardizing variables (often by subtracting the mean and dividing by the standard deviation) **does not fundamentally change the interpretation** of the coefficient in terms of its effect on the dependent variable for a one-unit change in the independent variable. However, it **facilitates comparison between coefficients or variables on a common scale and aids in assessing the relative importance** of predictors in the model.

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	69.609	8.357		8.329	.000	52.880	86.337
	RI[a]	-.427	.818	-.090	-.522	.604	-2.064	1.210
	Religiosity[a]	.265	.199	.237	1.334	.187	-.133	.663
	EA[a]	.188	.244	.098	.769	.445	-.301	.676
	PF[a]	.610	.209	.374	2.915	.005	.191	1.029

b. Predictors: (Constant), Religious involvement (RI), Religiosity, Emotion-avoidance coping (EA), and Problem-focused coping (PF).
pendent Variable: Challenges scale

Logistic regression

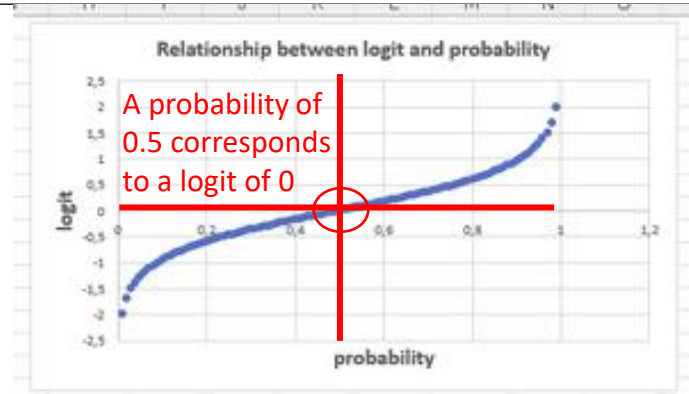
Probabilities

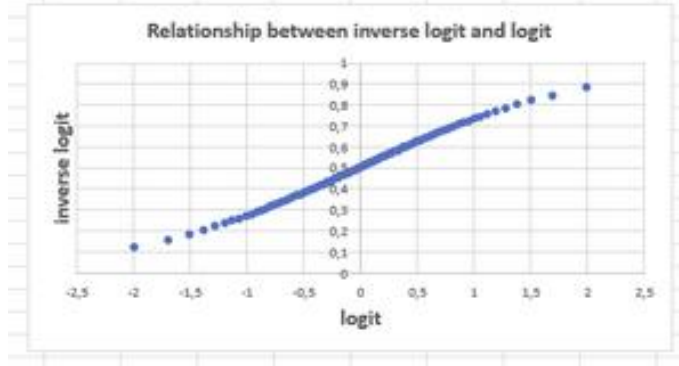
In logistic regression, Y must be transformed into p/(1-p) so that the binary dependent variable can be expressed as a function of continuous positive values ranging from 0 to unlimited (+). In this explanation, how can the probability go to positive unlimited?

In logistic regression, the goal is to **model the probability (p) of the occurrence of an event** (binary outcome) as a function of predictor variables. The logistic regression model assumes a relationship between the log-odds of the event and the predictor variables. The transformation, where the dependent variable Y is expressed as the **odds ratio (p/(1-p))**, is related to how logistic regression handles binary outcomes (e.g. absence or presence of an event).

Reminder:

Instead of modelling Y directly, logistic regression models the **log-odds (or logit)** of the event happening (p/(1-p)) as a linear combination of predictor variables. The inverse of the logit function is used to convert the predicted log-odds back into probabilities (p) between 0 and 1.





Odds ratio (OR)

Can we only interpret the sign and not the magnitude in odds ratio?

OR varies between 0 and +∞ (infinity) and is a measure used in logistic regression to quantify the relationship between the probability of an event occurring (e.g., winning a game) and the presence of a certain predictor.

The interpretation of the **OR** involves considering the **magnitude** of the ratio, contrary to the interpretation of the **log odds** (or logit) where only the **sign** is interpreted.

Example: if the odds ratio for game 1 is 0.2 and for game 2 is 0.1, the odds of winning in game 1 are indeed 2 times higher compared to game 2.

Logit(p) = log(p/(1-p)) = β0 + β1X1 + β2X2 + ⋯ + βnXn
The coefficients (e.g., β1) represent the change in log odds for a one-unit change in the corresponding predictor variable while holding other variables constant.

	Predictor	β	SEβ	Wald's χ^2	df	*p*	eβ (Odds Ratio)
Gender	female (v. male)	0.286	0.185	2.401	1	0.121	1.331
Education	uni (v. high school)	−0.490	0.231	4.486	1	0.034 *	0.613
	uni (v. tafe)	0.087	0.227	0.145	1	0.703	1.091
	tafe (v. high school)	−0.576	0.269	4.595	1	0.032 *	0.562
Age	18–34 years (v. 35–44 years)	−0.623	0.315	3.914	1	0.048 *	0.537
	18–34 years (v. 45–54 years)	0.266	0.261	1.042	1	0.307	1.305
	18–34 years (v. 55 yr or more)	0.039	0.242	0.027	1	0.871	1.040
	35–44 years (v. 45–54 years)	0.623	0.315	3.914	1	0.048 *	1.864
	35–44 years (v. 55 years or more)	0.396	0.305	1.679	1	0.195	1.485
	45–54 years (v. 55 years or more)	−0.227	0.263	0.744	1	0.388	0.797
Employment	Full Time (v. part time)	0.788	0.261	9.087	1	0.003 *	2.199
	Full Time (v. casual)	−0.228	0.412	0.305	1	0.581	0.796
	Full Time (v. not working)	0.486	0.229	4.492	1	0.034 *	1.626
	Part Time (v. casual)	−1.016	0.431	5.548	1	0.019 *	0.362
	Part Time (v. not working)	−0.302	0.265	1.300	1	0.254	0.739
	Casual (v. not working)	0.714	0.413	2.988	1	0.084	2.042
Settlement	City (v. rural)	−0.081	0.231	0.122	1	0.727	0.923
	City (v. town)	−0.050	0.225	0.050	1	0.823	0.951
	Town (v. rural)	−0.030	0.274	0.012	1	0.912	0.970

* Significant, $p < 0.05$.

OR is calculated as the exponentiation of the coefficient associated with a predictor: $OR = exp^{\beta}$
Exponentiating the log odds yields the OR, providing an interpretation of the effect size in terms of the relative change in odds associated with a unit change in the predictor.

Marginal predictions

Dichotomous covariate with marginal predictions: why does the marginal effect equate the difference in the adjusted predictions for two groups?

Marginal predictions provide the expected or average value of the dependent variable (outcome) for specific values or levels of the predictor variables, while averaging over the other predictors in the model.

- **Dichotomous covariate** (e.g., a binary variable representing two groups coded as 0 and 1): the **marginal effect is essentially the difference in the adjusted predictions for the two groups** defined by the dichotomous variable. This occurs because when we calculate the adjusted predictions for each group (0 versus 1), the marginal effect measures the change in the predicted outcome between these two groups.
- **Continuous covariate:** computing margins "as X changes from 0 to 1" refers to the interpretation of the marginal effect of that continuous variable. It is the average change in the predicted probability (P(Y=1)) for a one-unit change in the continuous predictor variable X, controlling for the other variables in the model.

Relationship between probability (p), odds, and log odds

If p=0.5 of success, odds=1 and log odds: 0, which suggests no effect

If p=0.8 of success, odds=4 and log odds: 1.38, which suggests a positive effect

If p=0.8 of failure (0.2 of success), odds=1/4 and log odds: -1.38, which suggests a negative effect

What we model. What we are interested in.

	Proportion (p)	Odds (for to against)	Odds (p/1-p)	Logit (Log_e odds) (Log_e p/1-p)	Logit to odds e^x	Logit to proportion (antilogit) $\frac{e^x}{(1+e^x)}$
A	5 out of 10 (0.5)	5 to 5	$\frac{0.5}{1-0.5}=1$	0	$e^0=1.0$	$\frac{e^0}{(1+e^0)}=0.5$
B	6 out of 10 (0.6)	6 to 4	$\frac{0.6}{1-0.6}=1.5$	0.41	$e^{0.41}=1.5$	$\frac{e^{0.41}}{(1+e^{0.41})}=0.6$
C	8 out of 10 (0.8)	8 to 2	$\frac{0.8}{1-0.8}=4.0$	1.39	$e^{1.39}=4.0$	$\frac{e^{1.39}}{(1+e^{1.39})}=0.8$
D	1 out of 10 (0.1)	1 to 9	$\frac{0.1}{1-0.1}=0.1$	-2.20	$e^{-2.20}=0.1$	$\frac{e^{-2.20}}{(1+e^{-2.20})}=0.1$

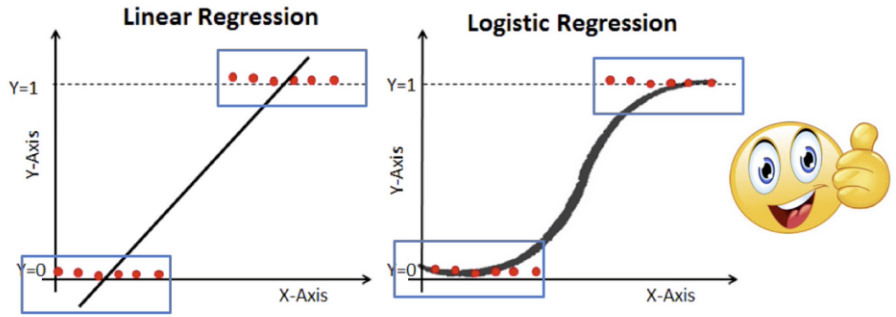
Linear Probability Model (LPM)

Is the Linear Probability Model the same as Logistic Regression or what exactly is the difference?

LPM is a basic regression model used to predict binary outcomes by applying linear regression on a binary response variable. In LPM, the dependent variable is directly modelled as a linear function of predictor variables. It **assumes a linear relationship** between predictors and the probability of a binary outcome. LPM **does not adhere to the constraints of probability**; therefore, predicted probabilities may fall outside the valid range of 0 to 1.

Logistic Regression is a type of regression **used specifically for binary outcomes**, employing the logistic function (sigmoid curve) to model the relationship between predictors and the log-odds of the outcome. It models the probability of a binary outcome as a **nonlinear function** of predictor variables, ensuring predicted probabilities lie between 0 and 1, as required for probabilities.

Linear Regression vs. Logistic Regression



Logistic Predicting probabilities that are between 0 and 1

ANOVA and repeated measures (RM) ANOVA

One-way ANOVA test

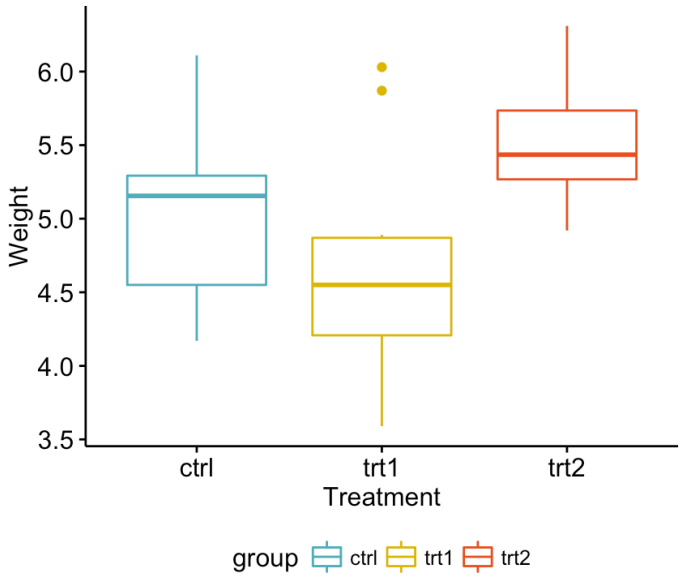
You have conducted the One-way ANOVA repeated measures test, and results indicate an F value of 29.34 with a Sig. (significance) level of .506. Does it suggest that the means are equal?

The overall F test indicates the means are equal. With a high p-value of .506, it suggests that there is insufficient evidence to reject the null hypothesis which assumes no differences among group means.

Further details (see next slide):

In one-way ANOVA test, a significant p-value indicates that some of the group means are different, but we don't know which pairs of groups are different.

It is possible to perform multiple pairwise-comparison (using Tukey Honest Significant Differences), to determine if the mean difference between specific pairs of group are statistically significant.



Post-hoc tests

```
> TukeyHSD(aov(lm(Slope~Condition,data=sshillNew)), ordered = TRUE)
  Tukey multiple comparisons of means
    95% family-wise confidence level
    factor levels have been ordered

Fit: aov(formula = lm(Slope ~ Condition, data = sshillNew))

$Condition
                     diff        lwr       upr     p adj
Neutral-Friends  10.47485  8.7150206 12.234679 0.0000000
Alone-Friends    10.83195  9.0721206 12.591779 0.0000000
Disliked-Friends 13.37140 11.6115706 15.131229 0.0000000
Alone-Neutral     0.35710 -1.4027294  2.116929 0.9507570
Disliked-Neutral  2.89655  1.1367206  4.656379 0.0002648
Disliked-Alone    2.53945  0.7796206  4.299279 0.0016721
```

What do levels mean in the context of ANOVAs?

In the context of ANOVA (Analysis of Variance), **"levels" refer to the different categories or groups within a factor**. When a factor has more than two levels (also called treatment groups), it means there are multiple categories or conditions being compared within that factor.

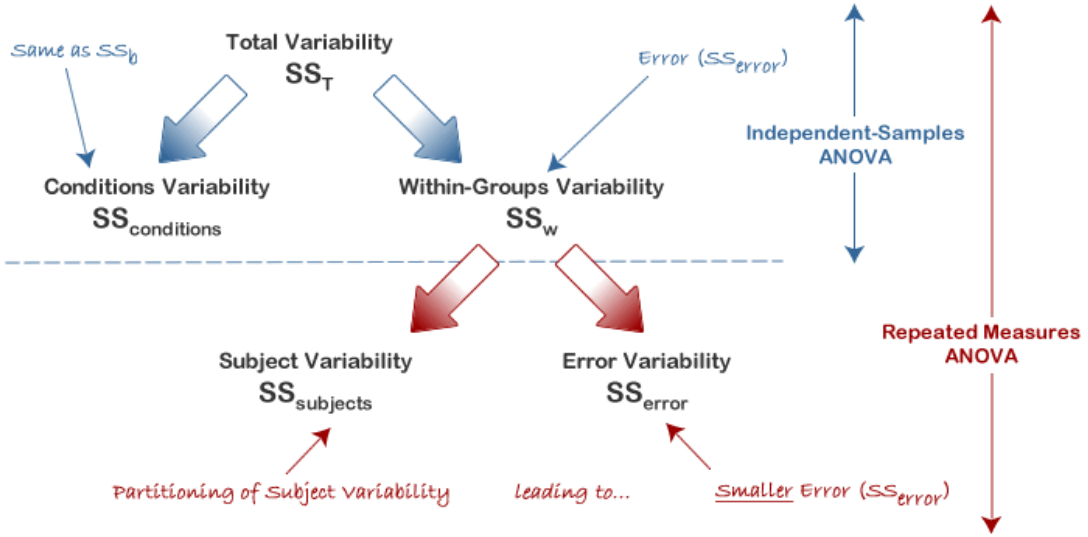
Why are Post hoc tests useful when a factor has more than two levels?

- ANOVA determines whether there are significant differences among the means of the groups, but it does not specify **which groups differ from each other**. Post hoc tests, conducted after finding a significant overall effect in ANOVA, help identify which specific pairs of groups differ significantly from each other.
- By identifying specific group differences, post hoc tests provide **more interpretable results**, allowing researchers to make more precise and informed conclusions about which groups are significantly different from each other.

F-statistic in RM ANOVA

How do individual differences affect the F-statistic in repeated measures ANOVA?

In repeated measures ANOVA, the individual difference portion of the error term is removed. Therefore, the denominator in the F-ratio will be smaller and the F will be larger. This means that the procedure will be more sensitive to small differences between groups.



$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

Example: When individual differences are substantial, the differences between groups might appear more pronounced, leading to a larger numerator (between-group variance) in the F-ratio relative to the denominator (within-group variance), resulting in a larger F-statistic, and potentially increasing the likelihood of obtaining a statistically significant F-statistic, especially with larger sample sizes.

Sphericity

Why is there an assumption that the variance of the differences between group should be equal (sphericity assumption)?

In repeated measures ANOVA, the assumption of sphericity refers to the **equality of variances among the differences between all possible pairs of conditions or levels within a repeated measures design**.

While the primary focus of repeated measures ANOVA is indeed to examine how groups (or conditions) differ from each other over time or under different conditions, the **sphericity assumption is related to the way the analysis considers the within-subjects variability**.

```
res <- anova_test(data = selfesteem, dv = score, wid = id, within
= time)
res
```

```
## ANOVA Table (type III tests)
##
## $ANOVA
##   Effect DFn DFd    F        p p<.05   ges
## 1   time   2  18 55.5 2.01e-08     * 0.829
##
## $`Mauchly's Test for Sphericity`
##   Effect     W     p p<.05
## 1   time 0.551 0.092
##
## $`Sphericity Corrections`
##   Effect  GGe      DF[GG]    p[GG] p[GG]<.05   HFe      DF[HF]
p[HF] p[HF]<.05
## 1   time 0.69 1.38, 12.42 2.16e-06         * 0.774 1.55, 13.94
6.03e-07         *
```

ges = generalized eta squared

significant p-value (p <= 0.05) indicates that the variances of group differences are not equal

When the assumption of sphericity is violated in repeated measures ANOVA, several **corrective measures** can be applied to address this issue, such as the **Greenhouse-Geisser and the Huynh-Feldt corrections**.

Mauchly's test of sphericity

When we assume that we have 3 factors in our RM ANOVA and Mauchly's test is significant ($p < 0.05$), does it mean that at least two variances of differences are unequal, or all variances of differences are unequal?

A **significant Mauchly's test** in a repeated measures ANOVA with three or more factors implies that the **assumption of sphericity is violated**, indicating that **at least one pair of within-subject conditions has unequal variances of the differences**. However, it **doesn't specify which pairs** of conditions specifically exhibit these unequal variances.

Moderation and mediation analyses

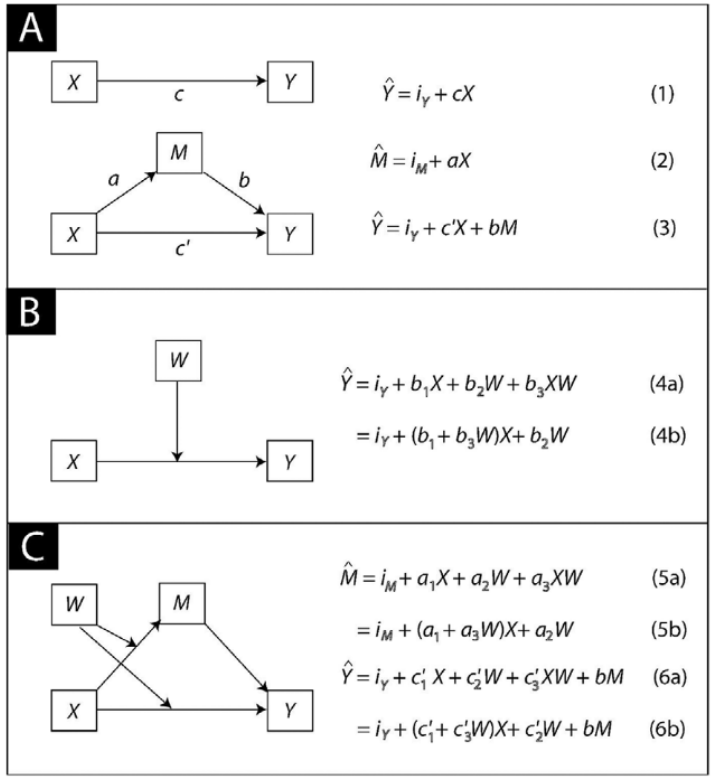


Figure 1. Simple Mediation (A), Simple Moderation (B), and a Conditional Process Model (C)

Igartua, J. J., & Hayes, A. F. (2021). Mediation, moderation, and conditional process analysis: Concepts, computations, and some common confusions. *The Spanish Journal of Psychology*, 24, e49.

Effects in moderation (and more generally)

What is the difference between main effect, simple main effect, simple effects, and conditional effects?

- A **main effects** refers to the effects of individual predictor variables (factors) on the dependent variable, ignoring the effects of other variables.

  $ME = \frac{1}{N} + \sum_{i=1}^{N} Y_i$, where Yi is the observed values of the dependent variable for each level of the factor, and N is the total number of observations.

- A **simple main effect** is the effect of an independent variable on the dependent variable at a specific level or combination of levels of another variable.

- **Simple effects** (similar to simple main effects) is the effect of an independent variable on the dependent variable separately for each level of another independent variable.

  $\mathrm{SE} = \frac{1}{n} + \sum_{i=1}^{n} Y_i$, where Yi is the observed values of the dependent variable for a specific level of Factor A at a given level of Factor B, and n is the number of observations at that specific combination of levels.

- **Conditional effects** refer to the effects of one variable on the dependent variable when the relationship between the variables is moderated.

  $CE = \beta_1 + \beta_2 X_2$,where β1 represents the intercept, β2 represents the coefficient for the predictor variable X2, and CE represents the conditional effect of X2 on the outcome variable.

Mean centering in moderation

When we mean center, which variables do we mean center?

When applying mean centering variables in moderation analysis, you typically center the predictor variables involved in the interaction term, specifically the continuous variables that are part of the interaction.

In moderation analysis, centering is often used to reduce multicollinearity and aid in the interpretation of interaction effects.

Conditional effect in moderation

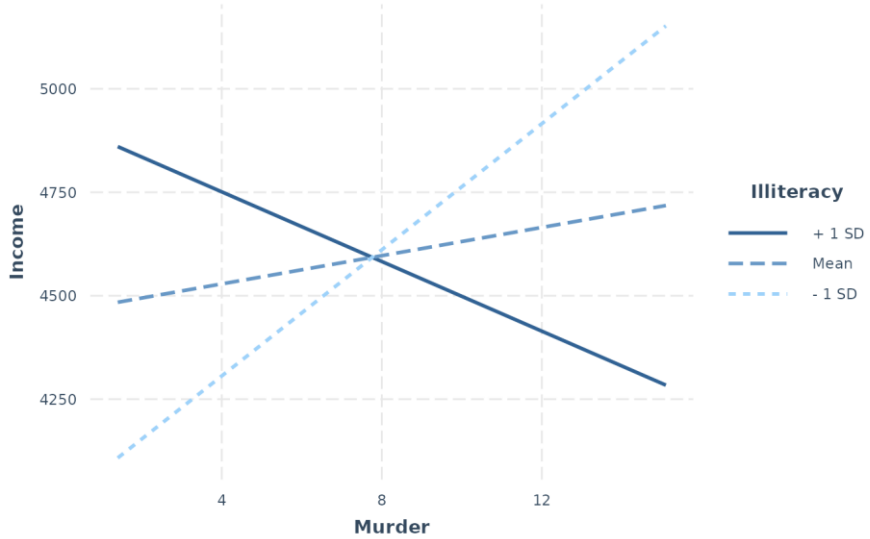
Simple slope / conditional effect of x - which part of the regression equation tells us this? Is it a b coefficient?

The simple slope or conditional effect of X in moderation analysis refers to the **effect of the independent variable (X) on the dependent variable (Y) at specific values or levels of the moderator variable (Z)**. This effect is typically examined within the context of a moderation model that includes interaction terms between X and Z. This interaction captures the effect of X on Y that varies across different levels of Z.

$Y = b0 + b1X + b2Z + b3(X*Z) + \varepsilon$

The conditional effect of X at a specific level of Z can be calculated by adding the coefficients of X (main effect) and the interaction term (X*Z) for that specific level of Z.

Example: the simple slope of X at the mean of Z might be calculated as: b1 + b3(mean of Z)



Mediation: focus on ab

But the total effect is calculated by c = c' + ab which means that ab = c-c'. Then, how to interpret the following statements: 1) size of ab is not determined by c or c'; 2) ab could be large even though c is small.

The size of ab is not determined by c, c', or their statistical significance or lackthereof. Statistical inference about mediation focuses on inference about the product of a and b. It is the difference between the total and direct effects of X.

We want to know whether zero can be plausibly discounted as a value of the indirect effect.

Given that the total effect is the sum of two effects that may have different signs, ab could be large even though c is small.

And if ab and c' are the same size but different in sign, c = 0.

This constitutes one reason why we should not insist that the total effect of X is statistically different from zero prior to conducting a mediation analysis.

Effect size in mediation

How do we measure effect size in mediation?

In mediation analysis, we want to compare the magnitudes of the indirect effect, direct effect, and total effect to assess the relative contribution of each.

To measure the **effect size of mediation**, we can rely on the ratio of the indirect to total effect, indicating that the **mediator accounts for a certain percentage of shared variance between the predictor and the outcome variable**. Furthermore, **R-squared** assesses variance accounted for in the model.

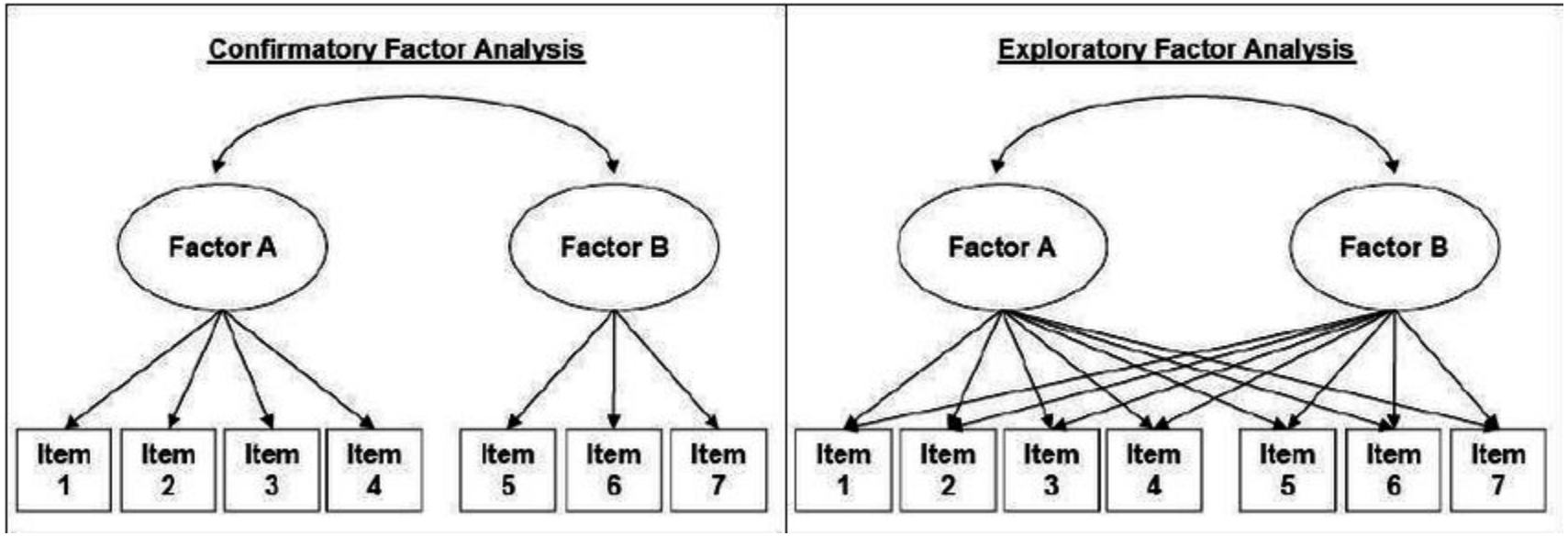
Reminder:

- **total effect (c):** relationship between X and Y, considering all paths (direct and indirect) without Z
- **indirect effect (ab):** relationship between X and Y through Z
- **direct effect (c'):** effect of X on Y after controlling for Z.

```
# define the model
modell.complete = "
## direct effect
liking ~ c*protest
## mediation path
respappr ~ a*protest
liking ~ b*respappr
## indirect effect (a*b)
ab := a*b
## total effect (c+a*b)
total := c+a*b
"
# get the complete output
fit.complete = lavaan::sem(modell.complete, data=sel)
lavaan::parameterestimates(fit.complete, standardized=T)[1:8]
##        lhs op      rhs label    est    se      z pvalue
## 1   liking  ~  protest     c -0.101 0.198 -0.508  0.611
## 2 respappr  ~  protest     a  1.440 0.220  6.544  0.000
## 3   liking  ~ respappr     b  0.402 0.069  5.857  0.000
## 4   liking ~~   liking        0.824 0.103  8.031  0.000
## 5 respappr ~~ respappr        1.354 0.169  8.031  0.000
## 6  protest ~~  protest        0.217 0.000     NA     NA
## 7       ab :=      a*b    ab  0.579 0.133  4.364  0.000
## 8    total :=    c+a*b total  0.479 0.193  2.478  0.013
lavaan::inspect(fit.complete,"r2")
##   liking respappr
##    0.246    0.249
```

Note about the code: better to specify cp (for c') instead of c.

Exploratory factor analysis (EFA)



Factor loadings

"In oblique rotation, element of a factor pattern matrix (= factor loading) is the unique contribution of the factor to the item". Isn't this statement wrong because in oblique rotation, it is the unique AND shared/common contribution?

The **factor pattern matrix** in oblique rotation emphasizes the **unique contribution of a factor to an item**. It is suggested that the elements in the factor pattern matrix predominantly represent the variance of an item that is uniquely explained by a specific factor and not shared with other factors.

Variables	Pattern Matrix		Structure Matrix	
	F_1	F_2	F_1	F_2
X_1	.800	.000	.800	.320
X_2	.700	.000	.700	.280
X_3	.600	.000	.600	.240
X_4	.000	.700	.280	.700
X_5	.000	.600	.240	.600
X_6	.000	.500	.200	.500

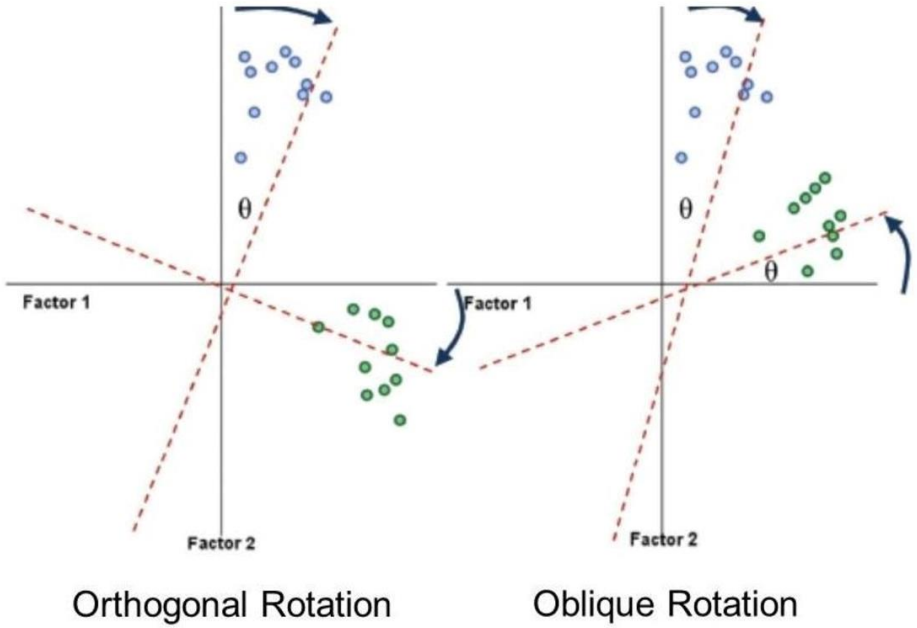
a. The pattern matrix is an exact replica of loadings shown in Figure 16. The structure matrix shows the correlations between the factors and variables.

The **factor structure matrix** focuses on the va**riance of an item that is not uniquely attributed to a single factor but rather shared or common across multiple factors**. The elements in the factor structure matrix represent the portion of an item's variance that overlaps or is explained by multiple factors, indicating the non-unique or shared contribution of factors to an item.

Oblique versus orthogonal rotation

Reminder:

With oblique rotation methods, factor loadings capture both unique and shared contributions due to allowed factor correlations. With orthogonal rotation methods, factor loadings tend to emphasize unique contributions and minimize the extent to which factors overlap in explaining the variance of the items.



Confirmatory factor analysis (CFA)

Mathematical requirements of CFA

What are "mathematical requirements". What does it mean?

In CFA, a mathematical requirement involves **specifying a measurement model that aligns with the underlying theoretical assumptions and meets certain conditions for the model to be identified**.

In the single choice question, why is the statement "Indicators should not correlate with one another" considered a mathematical requirement? Especially, because the statement "Items that should theoretically load on a factor should not correlate empirically" is not considered a mathematical requirement.

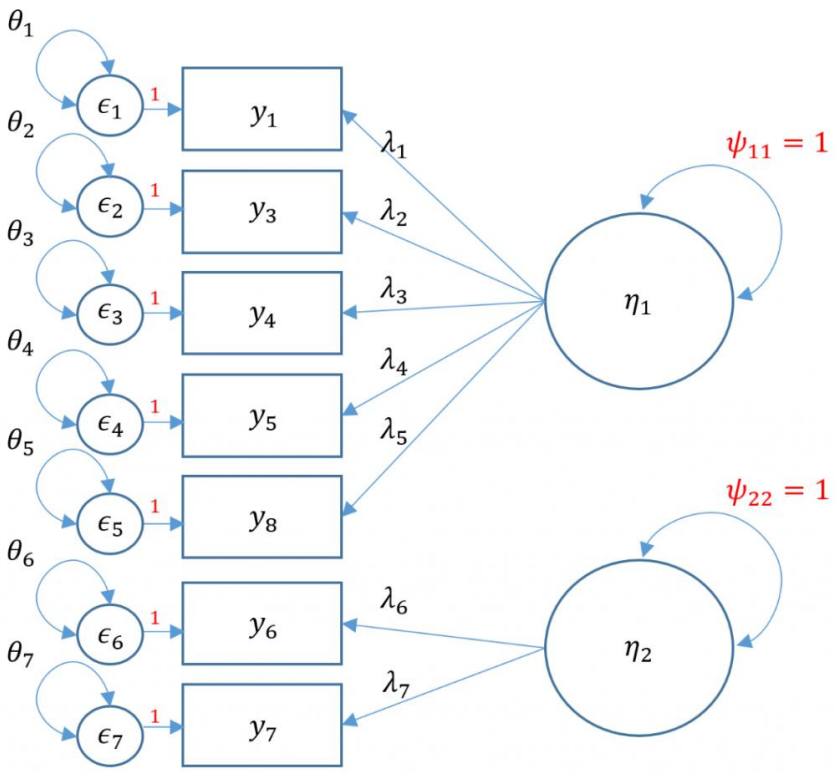
- In traditional CFA (especially when using orthogonal factors), the assumption or recommendation is often made that **indicators measuring different latent factors should not exhibit substantial correlations with each other**. This condition is imposed to ensure model identification and prevent multicollinearity issues, aiming for more straightforward interpretation and estimation of factor loadings.
- In contrast, the statement "Items that should theoretically load on a factor should not correlate empirically" is not true in the context of CFA since **we want items measuring the same latent construct or factor to show higher correlations with each other** due to their shared conceptual meaning or common underlying construct.

Constrained versus fixed parameters

What is the difference in CFA between constrained and fixed parameters, could you give an easy example for that?

Both constrained and fixed parameters serve to incorporate theoretical assumptions or restrictions into the CFA model.

- **constrained** parameters are estimated from the data under specific conditions (e.g., setting certain factor loadings to be equal)
- **fixed** parameters remain constant and are not estimated but rather imposed as predetermined values in the model



Example: for identification of the two-indicator factor model, we constrain the loadings of Item 6 and Item 7 to be equal, which frees up a parameter.

Model Chi-2 (in CFA)

Why is this statement true? The larger the model chi-square test statistic, the larger the residual covariance.

The model chi-square test **evaluates whether the observed data's covariance matrix fits the proposed model's covariance structure**.

A smaller chi-square value suggests a closer fit between the model and the observed data, indicating a better model fit.

More details (optional):

Since the model chi-square is proportional to the discrepancy of S (sample covariance matrix) and $\sum(\hat{\theta})$ (model implied matrix), the higher the chi-square the more positive the value of S - $\sum(\hat{\theta})$, defined as residual covariance.

$\sum(\hat{\theta})$ has the same dimensions as Σ, which is the observed population covariance matrix (matrix of bivariate covariances that determines how many total parameters can be estimated in the model).

Identification in CFA

Why is it important to ensure identification of parameters when conducting CFA?

Overall, ensuring identification in CFA is essential for obtaining reliable and interpretable results, facilitating accurate model estimation, assessing model fit, and deriving meaningful conclusions about relationships between latent constructs and observed variables.

More details:

To ensure identification in CFA, several strategies can be employed:

- Specifying a sufficient number of indicators for each latent construct.
- Employing identification constraints (e.g., fixing factor loadings, covariances, or intercepts) based on theoretical or empirical justifications.
- Using reference indicators or fixing factor loadings to set the scale or to identify the model.

Structural equation modelling (SEM)

SEM "balance"

"The SEM seeks to find a balance between maximizing the variance between constructs while ensuring the model remains parsimonious and generalizable."

What does maximizing the variance between constructs mean exactly?

Maximizing the variance between constructs refers to the idea that the latent constructs (factors or variables that are not directly observed but inferred from observed variables) should adequately explain the variance in the observed indicators that they represent.

And why does a higher variance between constructs often result in more precise estimates of the relationships among constructs?

Precision of estimates: when latent constructs effectively capture the variance in their indicators, it leads to more reliable and precise estimates of relationships among constructs in the model.

Parsimonious and generalizable models: SEM also aims for parsimony (having a simple model that adequately explains the observed data without unnecessary complexity), thus avoiding being overly complex or specific to the sample at hand. This, in turn, allows for broader applicability to other populations or settings.

Multilevel modelling (MLM)

Why using MLM?

If you have no reasons to expect the effect of your level-1 predictor to vary between clusters, the random intercept model is more appropriate than a random slope model. But when we do not expect an effect between clusters, why are we even performing an MLM?

Performing MLM, even when not expecting a meaningful effect between clusters, can still be beneficial for several reasons:

- **Accounting for nested structure**: even if you do not expect a substantial variation in the effects between clusters (no significant random slope), the data might still exhibit clustering or dependency within the clusters.
- **Adjustment for clustering effects**: ignoring the clustering structure in the data may violate assumptions of independence, leading to biased standard errors and potentially incorrect conclusions about the significance of the fixed effects.

  Example: random intercept models, explicitly model the variance at different levels (within-cluster and between-cluster variance), providing more accurate estimates of standard errors and inference.
- **Estimating variability at different levels**: understanding the between-cluster variability can provide insights into the overall differences or similarities among clusters, even if the focus is primarily on the fixed effects at the individual level.
- **Model comparison**: comparing different models helps in selecting the model that best fits the data.

Two-level null model

Imagine you have the following two-level null model specified for a dependent variable Yij (i: students and j: schools). What does the b0j represent? Why is it group-mean?

→ **b0j being the school-level intercept is considered the group-mean intercept** because it represents the average or typical value of the dependent variable within each school, allowing for variations across schools captured by the random effects **u0j**.

Level 1 equation: *Yij = b0j + rij*

- **Yij** denotes the observed dependent variable (e.g., student performance) for student i in school j.
- **b0j** represents the **intercept term for school j** (which is allowed to vary across schools, indicating that each school has its own intercept). It captures the **average outcome for students within a specific school**.
- **rij** represents the **residual or error term for each student i within school j** (individual-level variability not explained by the school-level intercept).

Level 2 equation: *b0j = g00 + u0j*

- **g00** is the **overall (or grand) intercept or average value of the dependent variable across all schools** (fixed part of the intercept common to all schools in the absence of random variability).
- **u0j** denotes the **random effect or school-specific deviation from the grand intercept** (variability across schools, allowing each school to have its own deviation from the overall average intercept g00).

Mean centering implication in MLM

Does the interpretation of b0j as the group-mean intercept suggest mean centering in the context of multilevel modeling?

Mean centering involves adjusting variables or model intercepts by subtracting a reference value (e.g. subtracting the overall grand mean g00), to facilitate interpretation or address issues like collinearity.

However, the explicit decision to use mean centering or the choice of the centering reference (e.g., grand mean or within-group mean centering) should be based on **theoretical considerations, statistical requirements, and the specific research context**, aiming to improve model interpretability or address certain statistical issues.

The interpretation of b0j as the group-mean intercept **might suggest an opportunity for mean centering** but does not directly imply the use of mean centering without further specification or justification based on the research objectives.

Bonus question

Bonus question example

Imagine that survey items investigating which factors are related to COVID-19 vaccine hesitancy for adults have been collected and that you are in charge of making a report. **Which method(s) of analysis would you use and why?**

Possible answer (not the only one!):

We would like to explain the propensity to hesitate about being vaccinated. Therefore, using a variable that measures respondents' willingness to be vaccinated (e.g., a 5-point scale), we could create a dichotomous dependent variable differentiating respondents willing to receive the vaccine (e.g., values 3-5 coded as '0') versus respondents not willing to get vaccinated (e.g., values 1-2 coded as '1'). We could then rely on logistic regression and include several predictor variables explaining what possible factors predict vaccine hesitancy (e.g., health condition, political leaning, education level, etc.).