Multilevel Data Analysis

AQ801-3-M & Version 1.1

Statistical Treatment of Clustered Data Part B

A·P·U
ASIA PACIFIC UNIVERSITY
OF TECHNOLOGY & INNOVATION

Contents & Structure

- Within-and between group relations: Regressions, Correlations, Estimation of within-and between-group correlations
- Combination of within-group evidence

Recap:

- What is the different between aggregation and disaggregation?

Learning Outcomes

- At the end of this topic, You should be able to:
  - Discuss the relationship between empty model and one way ANOVA
  - Discuss appropriate statistical methods for different multilevel data.

Key Terms You Must Be Able To Use

- If you have mastered this topic, you should be able to use the following terms correctly in your assignments and exams:

Within- and between-group regressions.

Combination of tests

Combination of estimates.

Overview of the Topic

- This part devoted to some statistical methods for multilevel data that attempt to uncover the role played by the various levels without fitting a full-blown hierarchical liner model.
- we describe the intraclass correlation coefficients, a basic measure for the between group effect in clustered observations.
- To avoid ecological fallacies it is essential to distinguish within-group from between-group regressions.

- These concepts are explained, and the relations are spelled out between within-group, between-group, and total regressions, and similarly for correlations.

Learning Outcome 1:

- **Discuss the relationship between empty model and one way ANOVA**

Empty Model

- In this section we assume a two-stage sampling design and infinite populations at either level.
- This macro-units will also be referred to as groups.
- A relevant model here is the random effects ANOVA model (one-way ANOVA model or empty model)

Empty Model

- Denoting by Y_{ij} the outcome value observed for micro-unit *i* within macro-unit j, this model can be expressed as
- $Y_{ij} = \mu + U_j + R_{ij}$ (3.1)
- Where μ is the population grand mean, U_j is the specific effect on macro-unit j, and R_{ij} is the residual effect for micro-unit *i* within this macro-unit.

Empty Model

- In other words, macro-unit j has the "true-mean" $\mu + U_j$, and each measurement of a micro-unit within this macro-unit deviates from this true mean by some value R_{ij}.
- Units differ randomly from one another, which is reflected in the fact that U_j is a random variable and the name "random effects model".

Empty Model

- Some units have a high true mean, corresponding to a high value of U_j, others have a true mean close to average, and still others a low true mean.
- It is assumed that all variables are independent, the group effects U_j having population mean 0 and population variance τ^2 (the population between-group variance), and the residuals having mean 0 and variance σ^2 (the population within-group variance)

The intraclass correlation

- The degree of resemblance (similarity) between micro-units belonging to the same macro-unit can be expressed by the ***intraclass correlation coefficient***.
- The use of the term “class” is conventional here and refers to the macro-units in the classification system under consideration.
- There are, however, several definitions of this coefficient, depending on the assumptions about the sampling design.

The intraclass correlation

- For example, if micro-units are pupils and macro-units are schools, then the within-group variance is the variance within the schools about their true means, while the between-group variance is the variance between the schools' true means.
- The total variance of Y_{ij} is then equal to the sum of these two variances,
- $var(Y_{ij}) = \tau^2 + \sigma^2$

The intraclass correlation

- The number of micro-units within the *j*th macro-unit is denoted by n_j.
- The number of macro-units is *N*, and the total sample size is $M = \sum_j n_j$.
- In this situation, the intraclass correlation coefficient ρ_I can be defined as
- $\rho_I = \dfrac{population\ variance\ between\ macro{-}units}{total\ variance}$
- $\rho_I = \dfrac{\tau^2}{\tau^2+\sigma^2}$ (3.2)

The intraclass correlation

- This is the proportion of variance that is accounted for by the group level.
- This parameter is called a correlation coefficient because it is equal to the correlation between values of two randomly drawn micro-units in the same, randomly drawn, macro-unit.
- Hedges and Hedberg (2007) report on a large variety of studies of educational performance in American schools, and find that values often range between 0.10 and 0.25

ICC: Testing for group differences

- The intraclass correlation as defined by (3.2) can be zero or positive.
- if it may be assumed that the within-group deviations R_{ij} are normally distributed, one can use an exact test for the hypothesis that the intraclass correlation is 0, which is the same as the null hypothesis that there are no group differences, or the true between-group variance is 0

ICC

- This is just the F-test for a group effect in the one-way analysis of variance (ANOVA).
- The test statistic can be written as

$$F = \frac{\tilde{n} S^2_{between}}{S^2_{within}}$$

- and it has an *F* distribution with *N*-1 and *M*-*N* degrees of freedom if the null hypothesis holds

Example 3.3 the F-test for the random data set

- For the data of Table 3.1, $F = \frac{10(105.7)}{789.7} = 1.34$ with 9 and 90 degrees of freedom.
- Thus, there is no evidence of true between-group differences.
- Statistical computer packages usually give the F-statistic and the within-group variances, S^2_{within}

- And the estimated intraclass correlation coefficient by
- $\hat{\rho}_I = \frac{F-1}{F+\tilde{n}-1} ---- (3.15)$

- If F<1, it is natural to replace both of these expressions by 0.
- These formulas show that a high value for the *F*-statistic will lead to large estimates for the between-group variance as well as the intraclass correlation.

- If there is not evidence for a main effect involving the group structure, then the researcher may leave aside the nesting structure and analyze the data by single-level methods such as ordinary least square (OLS) regression analysis.

Example: Random Data

- Suppose that we have a series of 100 observations as in the random digits in Table 3.1 in next slide.
- The core part of the table contains the random digits.
- Now suppose that each row in the table is a micro-unit, so that for each macro-unit we have observations on 10 micro-units.

Example

Table 3.1: Data grouped into macro-units (random digits from Glass and Stanley, 1970, p. 511).

j	Scores Y_{ij} for micro-units (random digits)										Average $\bar{Y}_{.j}$
01	60	36	59	46	53	35	07	53	39	49	43.7
02	83	79	94	24	02	56	62	33	44	42	51.9
03	32	96	00	74	05	36	40	98	32	32	44.5
04	19	32	25	38	45	57	62	05	26	06	31.5
05	11	22	09	47	47	07	39	93	74	08	35.7
06	31	75	15	72	60	68	98	00	53	39	51.1
07	88	49	29	93	82	14	45	40	45	04	48.9
08	30	93	44	77	44	07	48	18	38	28	42.7
09	22	88	84	88	93	27	49	99	87	48	68.5
10	78	21	21	69	93	35	90	29	12	86	53.4

Example

- The average of the scores for each micro-unit are in the last column.
- There seem to be large differences between the randomly constructed macro-units, if we look at the variance in the macro-unit averages (which is 105.65).
- The total observed variance between the 100 micro-units is 814.0.

Example

- Suppose the macro-units were schools, the micro-units pupils, and the random digits test scores.
- According to these two observed variances, we might conclude that the schools differ considerably with respect to their average test scores.

Learning Outcome 2:

- **Discuss appropriate statistical methods for different multilevel data.**

Within-group and between-group variance

- We continue to refer to the macro-units as groups.
- To understand the information contained in the data about the population between-group variance and the population within-group variance, we consider the *observed variance between groups* and the ***observed variance within-groups***.

Within-group and between-group variance

- The observed variance within group *j* is given by

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_{.j} \right)^2$$

- This number will vary from group to group.
- To have one parameter that expresses the within-group variability for all groups jointly, one uses the observed within-group variance.

Example 3.2: Within- and between-group variability for random data

- For the random digit table of the earlier example the observed between variance is $S^2_{between} = 105.7$

- The observed variance within the macro-units can be computed from formula (3.8)
- The observed total variance is known to be 814.0 and the observed between variance is given by 105.7

Example 3.2

- Then the estimated true variance within the macro-units is also $\hat{\sigma}^2 = 789.7$

Within- and between-group relations

- We saw in Section 3.1 that regressions at the macro level between aggregated variables $\bar{X}$ and $\bar{Y}$ can be completely different from the regressions between the micro-level variable X and Y.
- This section considers in more detail the interplay between macro-level and micro-level relations between two variables.

- focus is on regression of Y on X
- Total relations, that is relations at the micro-level when the clustering into macro-units is disaggregated, are mostly a kind of average of the within-group and between-group relations.

- Therefore it is necessary to consider within- and between-group relations jointly, whenever the clustering of micro-units in macro-units is meaningful for the phenomena being studied.

Regressions

- The linear regression of a “dependent” variable Y on an “explanatory” or “independent” variable X is the linear function of X that yields the best prediction of Y.
- When the bivariate distribution of (X,Y) is known and the data structure has only a single level, the expression for this regression function is
- $Y = \beta_0 + \beta_1 X + R$

- The constant term β_0 is called the intercept, while β_1 is called the regression coefficient.
- The term R is the residual or error component, and expresses the part of the dependent variable Y that cannot be approximated by a linear function of Y.

- Consider the artificial data set of Table 3.2 in the next slide.
- The first two columns in the table contain the identification numbers of the macro-units (*j*) and the micro-unit (*i*).
- The other four columns contain the data.
- X_{ij} is the variable observed for micro-unit *i* in macro-unit j, $\overline{X}_j$ the average of the X_{ij} values for group j, and similarly for the dependent variable *Y*.

Table 3.2: Artificial data on five macro-units, each with two micro-units.

j	i	X_{ij}	$\bar{X}_{.j}$	Y_{ij}	$\bar{Y}_{.j}$
1	1	1	2	5	6
1	2	3	2	7	6
2	1	2	3	4	5
2	2	4	3	6	5
3	1	3	4	3	4
3	2	5	4	5	4
4	1	4	5	2	3
4	2	6	5	4	3
5	1	5	6	1	2
5	2	7	6	3	2

- One might be interested in the relation between Y_{ij} and X_{ij}.
- The linear regression of Y_{ij} on X_{ij} at the micro level for the total group of 10 observations is
- $Y_{ij} = 5.33 - 0.33X_{ij} + R\ (total\ regression)$
- This is the disaggregated relation, since the nesting of micro-units in macro-units is **not** taken into account.

- The regression coefficient is -0.33.
- The aggregated relation is the linear regression relationship at the macro level of the group mean $\overline{Y_{.j}}$ on the group mean $\overline{X_{.j}}$. This regression line is
- $\overline{Y_{.j}} = 8.00 - 1.00\overline{X_{.j}} + R$

(regression between group means)

- The regression coefficient is now -1.00

Before running between group regression in *SAS Ent Guide*

- Tasks, Data, Query Builder
- Add a new computed Column, Summarized column, Language, Change to **AVG** and Put Column name (Avr_lang) REPEAT the same for Arithmetic (IQ)
- Untick “Automatically select group” Edit group, Group by schoolnr.

Use SAS Studio

- PROC SQL;
  CREATE TABLE MDA06768.MDA_DATA_AVG
  AS SELECT FLOOR(AVG(MDA_DATA.lang_post) *100)/100 AS Avg_lang, FLOOR(AVG(MDA_DATA.arit_post) *100)/100 AS Avg_arit FROM MDA06768.MDA_DATA
- GROUP BY schoolnr;

- A third option is to describe the relation between Y_{ij} and X_{ij} within each single group.
- Assuming that the regression coefficient has the same value in each group, this is the same as the regression of the within-group Y-deviations ($Y_{ij} - \overline{Y_{.j}}$) on the X-deviations ($X_{ij} - \overline{X_{.j}}$).

- This within-group regression line is given by
- $Y_{ij} = \overline{Y_{.j}} + 1.00\left(X_{ij} - \overline{X_{.j}}\right) + R$
  - (regression within groups) use ***SAS Uni Ed***
- With a regression coefficient of +1.00
- Finally (and this is how the artificial data set was constructed), Y_{ij} can be written as a function of the within-group and between group relations between Y and X.

Using SAS University Ed.

- *options validvarname=any;*
- *libname mda2 XLSX '/folders/myfolders/MDA/MDA2.XLSX';*
- *proc mixed data=mda2.sheet1;*
- *class schoolnr;*
- *model lang_dev= IQ_dev /solution;*
- *run;*

- This amounts to putting together the between-group and within-group regression equations.
- The result is
- $Y_{ij} = 8.00 - 1.00\bar{X}_{.j} + 1.00\left(X_{ij} - \overline{X_{.j}}\right) + R$
- $Y_{ij} = 8.00 + 1.00X_{ij} - 2.00\overline{X_{.j}} + R$
  - (multilevel regression) – use ***SAS Uni Ed***

- The five parallel ascending lines in the next slide represent the within-group relation between Y and X.
- The steep descending line represents the relation at the aggregate level (i.e. between the group means), whereas the almost horizontal descending line represents the total relationship, that is, the micro-level relation between X and Y ignoring the hierarchical structure.

Figure 3.4 graphically depicts the total, within-group, and between-group relations between the variables.

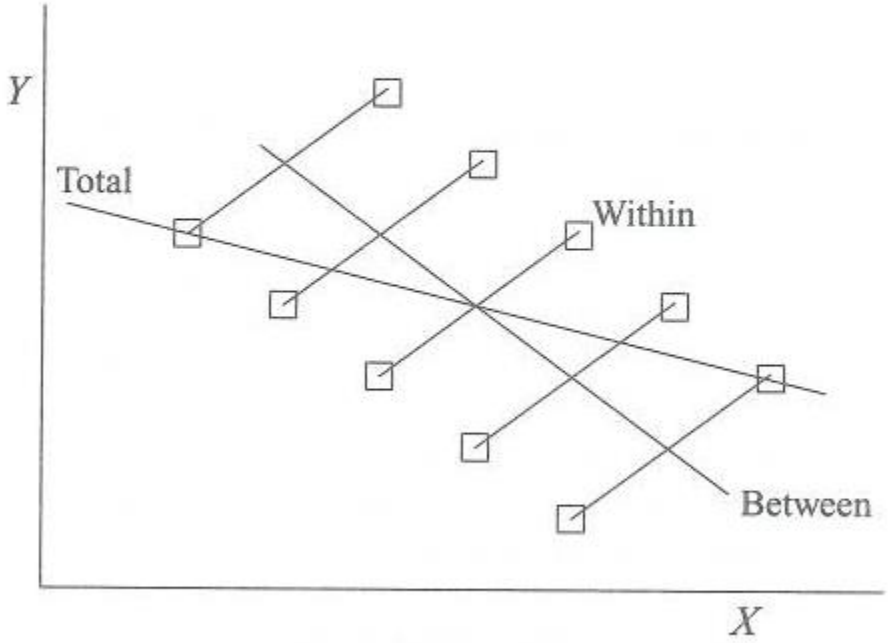

Figure 3.4: Within, between, and total relations.

- The within-group regression coefficient is +1 whereas the between-group coefficient is -1.
- The total regression coefficient, -0.33, lies in between these two.
- This illustrates that within-group and between-group relations can be completely different, even have opposite signs.

- The true relation between Y and X is revealed only when the within- and between-group relations are considered jointly, that is, by the multilevel regression.
- In the multilevel regression, both the between-group and within-group regression coefficients play a role.

Summary of Main Teaching Points

- Total regression did not consider the multilevel data structure, between group regression, within group regression and multilevel regression will be more efficient.

Question and Answer Session

Q & A

What we will cover next

- **The Random Intercept Model**