Regression modelling for Reliability/ICC in Stata

Niels Henrik Bruun

Research data and statistics, Aalborg University Hospital

1. Introduction
2. Intraclass correlation (ICC) / reliability
3. Challenges
4. Statistics today
5. Continuous measurements
6. Ordered or binary measurements
7. Summary

Section 1

Introduction

Questions asked regarding Reliability / ICC

- "Advanced techniques are possible for researchers who are interested in providing more information than a summary statistic", Hernaez (2015)
- Focus: Intraclass correlation (ICC)
  - most versatile and most potential
  - Is the classical black box framework the proper way today?
  - How does Stata support more modern approaches?
  - Code examples

Section 2

Intraclass correlation (ICC) / reliability

Definition of agreement, Vet et al. (2006), Hernaez (2015)

- Measurement agreement is *Measurement variation*
  - How fine can one measure?
  - A kitchen weight may weight correct within $\pm 5g$
- The level of non-dectechable variation due to instrument

Definition of reliability / ICC

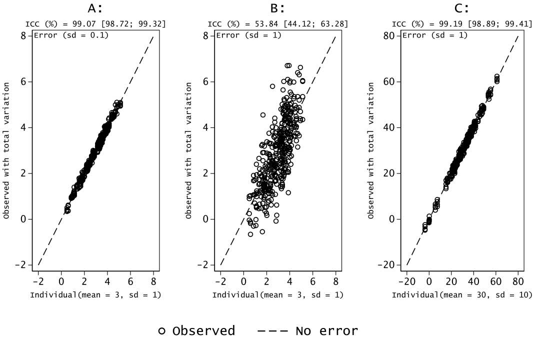
- How well measurements are distinguished despite *Measurement variation*
  - $1 - \textit{reliability}$ is the degree of bias due to *Measurement variation*
  - A bath weight (correct within $\pm 1kg$) is useless in a kitchen

- $\textit{reliability} = \frac{\textit{Variation between study objects}}{\textit{Variation between study objects} + \textit{Measurement variation}}$, Streiner, Norman, and Cairney (2015)

- $\textit{Variation} = \textit{Variance} \Rightarrow \textit{ANOVA}$?
- ICC ranges from 0 (no reliability) to 1 (perfect reliability)
- ICC correlates variables with the same class (unit) and variance, McGraw and Wong (1996)
  - in contrast to eg Pearsons correlation (example: height(cm) vs weight(kg))

- $\textit{Variation between study objects} / \textit{Measurement variation} = \textit{reliability} / (1 - \textit{reliability}))$
  - $\textit{reliability} = 0.5 \Rightarrow \textit{Variation between study objects} / \textit{Measurement variation} = 1$
  - $\textit{reliability} = 0.8 \Rightarrow \textit{Variation between study objects} / \textit{Measurement variation} = 4$
  - $\textit{reliability} = 0.9 \Rightarrow \textit{Variation between study objects} / \textit{Measurement variation} = 9$
- See Koo and Li (2016) for interpretation and reporting

Effect of agreement / *Measurement variation* on observed



- A and B: Same *Variation between study objects*, different *Measurement variation*
- A and C: same *Variation between study objects* relative to *Measurement variation*
- B and C: Different *Variation between study objects*, same *Measurement variation*
- See Dunn (1989) and Vet et al. (2011) on Generalisability theory and reliability

Textbook dataset layout

	measurement 1	measurement 2	...	measurement k
subject 1	y_{11}	y_{12}	...	y_{1k}
subject 2	y_{21}	y_{22}	...	y_{2k}
...	...	...	...	...
subject n	y_{n1}	y_{n2}	...	y_{nk}

- n subjects (rows) are having k measurements (columns)
- Measurements in cells are typically not repeated
- Balanced design of single values
- Possible bias from measurements (columns)
- Shrout and Fleiss (1979) and McGraw and Wong (1996) propose a **standardised** setup based on ANOVA
    - Continuous measurements

Section 3

Challenges

Is ANOVA the best starting point for ICCs today?

- Serious weaknesses of ANOVA estimators, Marchenko (2006)
  - Possibly negative estimates of variance components
  - Nonexistence of uniformly best estimators
  - Lack of uniqueness in the case of unbalanced data
- Shrout and Fleiss (1979) and McGraw and Wong (1996) made their suggestion in the early pc years
- How to handle ordered or categorical outcomes properly?
- Do some measurements needs adjustment?
  - Example: Measurement precision might dependent on age?

Research design and reliability, Zacho et al. (2020)

- Four raters from two hospitals using a standard and a new method
  - Two raters from each hospital
    - n subjects for each rater
    - All subjects are rated twice
    - All raters has used both methods on the n subjects
  - 3 months later a second rating
    - All subjects are rated once
    - Standard method is rated within hospital one, new method within hospital two
- Outcome has 3 levels:
  - Benign - 60%
  - In doubt - 20%
  - Malignant - 20%
- How many research questions are hidden behind this design?
  - Is a set of pairwise comparisons by ICC (or Kappa) the best way to analyze?

Section 4

Statistics today

On Anova, maximum likelihood (ML) and restricted maximum likelihood (REML)

Marchenko (2006) (also see Rabe-Hesketh and Skrondal (2012)):

- REML and ML variance estimates are guaranteed to be nonnegative
- REML takes into account the implicit degrees of freedom associated with the fixed effects
- ANOVA and REML estimators are identical for balanced designs
- For unbalanced designs, all three estimators generally differ
- ML and REML are preferred methods of estimation for unbalanced data due to simplicity

ICC simplified, Liljequist (2019)

Name	Model	ICC (agreement)
oneway	$y_{ij} = \mu + R_i + E_{ij}$	$\frac{\sigma_R^2}{\sigma_R^2+\sigma_E^2}$
twoway random	$y_{ij} = \mu + R_i + C_i + E_{ij}$	$\frac{\sigma_R^2}{\sigma_R^2+\sigma_C^2+\sigma_E^2}$
twoway fixed	$y_{ij} = \mu + R_i + c_i + E_{ij}$	$\frac{\sigma_R^2}{\sigma_R^2+\hat{\sigma}_c^2+\sigma_E^2}$

- Capital letters are random effects
- Interaction between subjects and measurements as part of the Error
- Same ICC formulas for twoway mixed (pseudo $\hat{\sigma}_c^2$) and twoway random
- Bias over measurements / columns
  - Agreement or Consistency, see McGraw and Wong (1996) p. 33
  - Agreement (same level?)
  - Consistency (Same order?): Leave out bias by measurements $\hat{\sigma}_c^2$ or σ_C^2
- Do three ICC formulas; oneway; twoway agreement; and twoway consistency

Section 5

Continuous measurements

PEFR example from Rabe-Hesketh and Skrondal (2012) or Bland and Altman (1986)

17 subjects have their peak expiratory flow rate (PEFR) measured twice with two different instrument

```
use "http://www.stata-press.com/data/mlmus3/pefr", clear
reshape long wp wm, i(id) j(time)
reshape long w, i(id time) j(pfmeter) string
rename w pefr
strtonum pfmeter
label define pfmeter 1 "mini Wright (l/min)" 2 "Wright (l/min)", replace
```

Using -icc-

- you cannot have repeated measurements in twoway -ICC-

```
icc pefr id pfmeter if time == 1

Intraclass correlations
Two-way random-effects model
Absolute agreement

Random effects: id                Number of targets =        17
Random effects: pfmeter           Number of raters  =         2

--------------------------------------------------------------
                  pefr |        ICC       [95% conf. interval]
-----------------------+--------------------------------------
            Individual |   .9459284       .8574112    .9800787
               Average |    .972213       .9232325    .9899391
--------------------------------------------------------------
F test that
  ICC=0.00: F(16.0, 16.0) = 34.03             Prob > F = 0.000

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

Using -mixed- and -nlcom-

- To get same ICCs as from -icc-, the variance components must be crossed
- Only one component needs to be crossed, see recipe in Marchenko (2006) and Rabe-Hesketh and Skrondal (2012)
- Confidence intervals not quite the same as for -icc-
- For comparison we only look at time 1

```
mixed pefr if time == 1, reml noheader nolog nofetable ||id: ||_all: R.pfmeter
nlcom ( icc_i: exp(2*_b[lns1_1_1:_cons]) / (exp(2*_b[lns1_1_1:_cons]) ///
  + exp(2*_b[lns2_1_1:_cons]) + exp(2*_b[lnsig_e:_cons])) ), noheader post

------------------------------------------------------------------------------
        pefr | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       icc_i |      0.946      0.026    36.56   0.00         0.895       0.997
------------------------------------------------------------------------------
```

Using -mixed- and -estat icc-

- -estat icc- do not work for crossed effects
- Described in eg Rabe-Hesketh and Skrondal (2012)
- Formula for confidence intervals, see StataCorp LLC (2021 ME) p. 55-56
- For comparison we only look at time 1

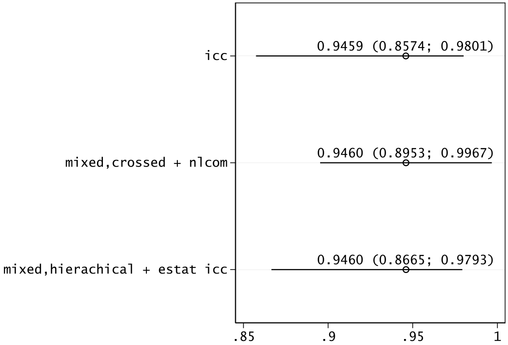
```
mixed pefr if time == 1, reml noheader nolog nofetable ||id: ||pfmeter:
estat icc

Intraclass correlation

------------------------------------------------------------------------------
                       Level |        ICC   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
                          id |   .9460141   .0258752      .8665189    .9792968
                  pfmeter|id |    .980654   3.482216      2.9e-155           1
------------------------------------------------------------------------------
```

Summary, Continuous measurements

- Similar ICC estimates
- Mixed with crossed variance confidence interval more similar to traditional ICC
- Both mixed-effect models are with option **reml**
- Several -gsem- attempts with no convergence (I'm no -gsem- expert)

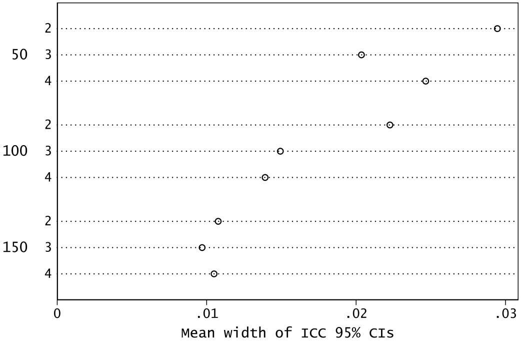


A note on power calculations, Continuous measurements

- Lew and Doros (2010) suggests simulations to find optimal for *n* and *k* wrt mean width of ICC 95% CI
- Mata and -simulate- makes it easy to simulate the datasets using the kronecker operator (#)
- Optimal solution for *n* and *k* depends on $\sigma_{subject}$, ($\sigma_{measurement}$) and agreement σ_{error}
  - Example code next slide for the values 1, 0.3 and 0.1 respectively
    - In this case more subjects is better
    - More raters is not necessarely better
- Alternative is to get the probability of ICC being above a chosen limit, eg 0.8
- On next slide
  - (n,k) = (50, 3) is better than (100, 2)
  - Precision between 0.01 and 0.03

Power simulation, Continuous measurements

```
capture program drop ciw1
program define ciw1, rclass
    args n k mu sd_s sd_m err
      clear
      mata: out = (1::`n') # J(`k', 1, 1)
      mata: out = out, rnormal(`n', 1, 0, `sd_s') # J(`k', 1, 1)
      mata: out = out, J(`n', 1, 1) # (1::`k')
      mata: out = out, J(`n', 1, 1) # rnormal(`k', 1, 0, `sd_m')
      mata: out = out, out[., 2] + out[., 4] + rnormal(`n'*`k',1,`mu',`err')
      mata: nhb_sae_addvars(("s", "m_s", "m", "m_m", "y"), out) // matrixtools
      mixed y, reml ||s: ||_all:R.m
      local icc_i_formula exp(2*_b[lns1_1_1:_cons])
      local icc_i_formula `icc_i_formula' / ( exp(2*_b[lns1_1_1:_cons])
      local icc_i_formula `icc_i_formula' + exp(2*_b[lns2_1_1:_cons])
      local icc_i_formula `icc_i_formula' + exp(2*_b[lnsig_e:_cons]) )
      if `e(converged)' {
        nlcom ( icc_i: `icc_i_formula'), post
        lincom _b[icc_i]
        return scalar ciw = r(ub) - r(lb)
      }
      else return scalar ciw = .
  end
forvalues n = 50(50)150 {
    forvalues k = 2/4 {
      quietly simulate ciw = r(ciw), reps(20) nodots: ciw1 `n' `k' 10 2 0.3 0.1
      quietly g n = `n'
      quietly g k = `k'
      quietly if !(`n' == 50 & `k' == 2) append using data/icc1
      quietly save data/icc1, replace
    }
  }
graph dot (mean) ciw, over(k) over(n) name(fig2, replace) ytitle(Mean width of ICC 95% CIs)
```



Section 6

Ordered or binary measurements

Rating example from StataCorp LLC (2021)

6 subjects (target) are measured by three different raters (judge) using a 1-10 scale (rating)

- ordered logistic regression (-meologit-) is often suggested when outcomes are scores
- -melogit- for binary measurements

```
use "https://www.stata-press.com/data/r17/judges", clear
```

Using -icc-

```
icc rating target judge

Intraclass correlations
Two-way random-effects model
Absolute agreement

Random effects: target          Number of targets =        6
Random effects: judge           Number of raters  =        4

--------------------------------------------------------------
               rating |        ICC       [95% conf. interval]
----------------------+---------------------------------------
           Individual |    .2897638       .0187865    .7610844
              Average |    .6200505       .0711368     .927232
--------------------------------------------------------------
F test that
  ICC=0.00: F(5.0, 15.0) = 11.03              Prob > F = 0.000

Note: ICCs estimate correlations between individual measurements
      and between average measurements made on the same target.
```

Using -meologit- and -nlcom-

- only one component needs to be crossed, see recipe in Marchenko (2006) and Rabe-Hesketh and Skrondal (2012)
- negative lower bound

```
meologit rating, noheader nolog ||_all: R.judge ||target:
nlcom ( icc_i: _b[var(_cons[target])] / (_b[var(_cons[target])] ///
  + _b[var(_cons[_all>judge])] + _pi^2/3) ), noheader post

------------------------------------------------------------------------------
      rating | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       icc_i |      0.302      0.191     1.58   0.11       -0.073       0.676
------------------------------------------------------------------------------
```

Using -meologit- and -estat icc-

- Error variance for a mixed-effects logistic and ordered logistic regression is $\pi^2/3$, StataCorp LLC (2021 ME) p. 55
- Option **intpoint(20)** is to achieve convergence

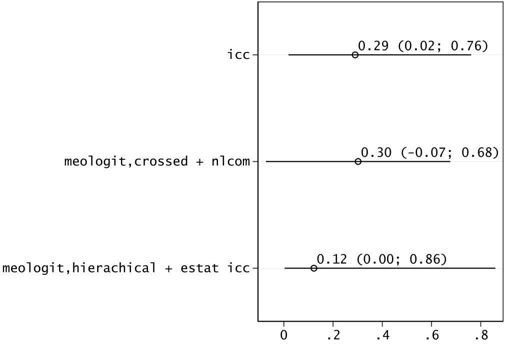
```
meologit rating, noheader nolog intpoint(20) ||target: ||judge:
estat icc

Residual intraclass correlation

------------------------------------------------------------------------------
                       Level |        ICC   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
                      target |   .1222287    .2069544      .0031659    .8592619
                judge|target |   .9349675    .5611143      2.01e-07           1
------------------------------------------------------------------------------
```

Summary, ordered measurements

- wide confidence intervals
- "meologit,crossed + nlcom" gives similar estimates to traditional ICC
- "meologit,crossed + nlcom" has a negative lower bound
- "meologit,hierachical + estat icc" gives quite a different estimate
- -gsem- not tested

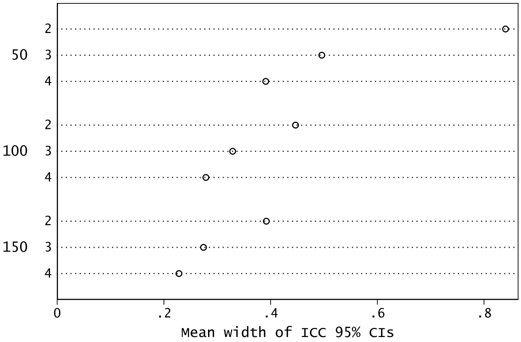


A note on power calculations, ordered or binary measurements

- Lew and Doros (2010) suggests simulations to find optimal for *n* and *k* wrt mean width of ICC 95% CI
- Mata and -simulate- makes it easy to simulate the datasets using the kronecker operator (#)
- Inspiration from Buis (2007) and Statalist, Xavier, 2021-05-17
- Use code next slide with **caution**, see Statalist, Enzmann, 2016-06-21
- Optimal input should include approximate distribution of the score
- Challenge: Interpretation of SDs in the random effects
- On next slide precision much lower (between 0.2 and 0.8)
    - Choose n and k as big as possible
    - Note (n,k) = (100,3) is better than (150,2)

Power simulation, ordered or binary measurements

```
capture program drop ciw2
program define ciw2, rclass
    args n k sd_s sd_m percs
      clear
      mata: out = (1::`n') # J(`k', 1, 1)
      mata: out = out, rnormal(`n', 1, 0, `sd_s') # J(`k', 1, 1)
      mata: out = out, J(`n', 1, 1) # (1::`k')
      mata: out = out, J(`n', 1, 1) # rnormal(`k', 1, 0, `sd_m')
      mata: out = out, out[., 2]+out[., 4]+rnormal(`n'*`k',1,0,pi()/sqrt(3))
      mata: nhb_sae_addvars(("s", "m_s", "m", "m_m", "xb"), out) //matrixtools
      g rlogit = logit(runiform())
      _pctile xb, percentiles(`percs')
      g y = (rlogit > xb + r(r1)) + (rlogit > xb + r(r2))
      capture meologit y ||_all:R.m ||s:
      if ! _rc {
        nlcom ( icc_i: _b[var(_cons[s])] / (_b[var(_cons[s])] ///          + _l
        lincom _b[icc_i]
        return scalar ciw = r(ub) - r(lb)
      }
      else return scalar ciw = .
  end
forvalues n = 50(50)150 {
    forvalues k = 2/4 {
      quietly simulate ciw = r(ciw), reps(20) nodots: ciw2 `n' `k' 2 0.3 "60 80"
      quietly g n = `n'
      quietly g k = `k'
      quietly if !(`n' == 50 & `k' == 2) append using data/icc2
      quietly save data/icc2, replace
    }
  }
graph dot (mean) ciw, over(k) over(n) name(fig3, replace) ytitle(Mean width of ICC 95% CIs)
```



Section 7

Summary

Take home

- From a statistical view, it is better to work modelbased
  - Model control
  - Transformations?
  - Unbalanced datasets
  - Use of designs
  - Power (simulation) calculations
- On effects (crossed vs hierachical)
  - StataCorp LLC (2021), and eg Rabe-Hesketh and Skrondal (2012) concentrates on ICC based on hierachical effects
  - ICC based on models with crossed effects more similar with ANOVA
  - In Stata -estat icc- only works with hierachical models
- Use -meologit-/-melogit- and $sd^2_{error} = \pi^2/3$ for ordered/binary categorical variables
  - Challenge: Interpretation of SDs in the random effects
- -gsem- should be appealing - more work required

Questions?

- **Thank you!!**
- References on next slide

References I

Bland, J. Martin, and Douglas G. Altman. 1986. "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement." *The Lancet* 327 (8476): 307–10. https://doi.org/https://doi.org/10.1016/S0140-6736(86)90837-8.

Buis, M. L. 2007. "Stata Tip 48: Discrete Uses for Uniform()." *Stata Journal* 7 (3): 434–435(2). //%3C?%20echo(www)%20?%3E.stata-journal.com/article.html?article=pr0032.

Dunn, G. 1989. *Design and Analysis of Reliability Studies*. Edward Arnold Publishers.

Hernaez, Ruben. 2015. "Reliability and Agreement Studies: A Guide for Clinical Investigators." *Gut* 64 (April). https://doi.org/10.1136/gutjnl-2014-308619.

Koo, Terry K., and Mae Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15 (2): 155–63. https://doi.org/https://doi.org/10.1016/j.jcm.2016.02.012.

Lew, Robert, and Gheorghe Doros. 2010. "Design Based on Intra-Class Correlation Coefficients." *Current Research in Biostatistics* 1 (1): 1–8. https://doi.org/10.3844/amjbsp.2010.1.8.

Liljequist, Britt AND Skavberg Roaldsen, David AND Elfving. 2019. "Intraclass Correlation – a Discussion and Demonstration of Basic Features." *PLOS ONE* 14 (7): 1–35. https://doi.org/10.1371/journal.pone.0219854.

Marchenko, Y. V. 2006. "Estimating Variance Components in Stata." *Stata Journal* 6 (1): 1–21(21). http://www.stata-journal.com/article.html?article=st0095.

McGraw, K. O., and S. P. Wong. 1996. "Forming Inferences About Some Intraclass Correlation Coefficients." *Psychological Methods* 1 (1): 30–46.

References II

Rabe-Hesketh, S., and A. Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata: Continuous Responses, Third Edition*. vb. 1. Stata Press.

Shrout, Patrick E, and Joseph L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86 2: 420–8.

StataCorp LLC, TX, College Station. 2021. "Stata 16 Base Reference Manual." https://www.stata.com.

Streiner, David L., Geoffrey R. Norman, and John Cairney. 2015. *Health Measurement Scalesa Practical Guide to Their Development and Use: A Practical Guide to Their Development and Use*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/med/9780199685219.001.0001.

Vet, Henrica C. W. de, Caroline B. Terwee, Lidwine B. Mokkink, and Dirk L. Knol. 2011. *Measurement in Medicine: A Practical Guide*. Practical Guides to Biostatistics and Epidemiology. Cambridge University Press. https://doi.org/10.1017/CBO9780511996214.

Vet, Henrica de, Caroline Terwee, Dirk Knol, and Lex Bouter. 2006. "When to Use Agreement Versus Reliability Measures." *Journal of Clinical Epidemiology* 59 (November): 1033–9. https://doi.org/10.1016/j.jclinepi.2005.10.015.

Zacho, Helle D., Ramune Aleksyniene, June A. Ejlersen, Joan Fledelius, and Lars J. Petersen. 2020. "Inter- and intraobserver agreement in standard and ultra-fast single-photon emission computed tomography/computed tomography for the assessment of bone metastases." *Nuclear Medicine Communications* 41 (10): 1005–9. https://doi.org/10.1097/MNM.0000000000001252.