Interrater Reliability in Healthcare Studies:

Calculating the Intraclass Correlation Coefficient (ICC) in SAS

Ellen Maki, Ph.D.
Analytica Statistical Consulting Inc.

Interrater Reliability

- Each subject assessed by multiple raters
- To what extent are the ratings within a subject homogeneous?
- Ideally, want raters to be interchangeable

Decayed, Missing, Filled Teeth

Patient	Examiner			
	1	2	3	4
1	8	7	11	7
2	13	11	15	13
3	0	0	2	1
4	3	6	9	6
5	13	13	17	10
6	19	23	27	18

J.L. Fleiss *The Design and Analysis of Clinical Experiments*

Intraclass Correlation Coefficient

- For continuous data, ICC often used to assess interrater reliability
- ICC is the correlation between two measurements made on same subject

$$ICC = \text{Corr}(Y_{ij}, Y_{ik})$$

ICC Properties

- Like any correlation, $-1 \leq ICC \leq 1$
- Ideally, $ICC \geq 0$
- ICC Values close to 1 are desirable and indicate good interrater reliability

About the ICC

"There are numerous versions of the intraclass coefficient (ICC)...Each form is appropriate for specific situations defined by the experimental design..."

Shrout, PE and Fleiss, JL (1979) "Intraclass Correlations: Uses in Assessing Rater Reliability", *Psychological Bulletin*, 86:2, 420-428

Randomized Block Design

- Dental example: 4 different raters randomly selected to rate each patient
- Patients represent random sample of all possible patients

$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$$

$\beta_i \sim N(0, \sigma_\beta^2)$ random subject effect

$\varepsilon_{ij} \sim N(0, \sigma^2)$ experimental error

ICC for Randomized Blocks

$$Cov(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik})$$
$$= \mu^2 + E(\beta_i^2) - \mu^2$$
$$= \sigma_\beta^2$$

$$Var(Y_{ij}) = Var(Y_{ij}) = \sigma_\beta^2 + \sigma^2$$

$$\therefore Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma^2} = ICC$$

Estimating the ICC: Method 1

- **PROC GLM**
  1. **Output mean squares to dataset**
  2. **Use mean squares to calculate estimates of σ_β^2 and σ^2**
  3. **Use the estimates of σ_β^2 and σ^2 to calculate the ICC**

Estimating the ICC: More Methods

- **PROC MIXED**
  1. Output estimates of variance components (part of standard output) to a dataset
  2. Use the estimates to calculate ICC
- **PROC NLMIXED**
  1. Calculate ICC within the procedure in a single step
- **%INTRACC macro**
  1. No programming to do!

Decayed, Missing, Filled Teeth

Start with PROC MIXED + PROC SQL Approach

```
ods output CovParms = cov1 ;
proc mixed data=dental method=ml;
      class patient ;
     model score =  ;
      random patient ;
run ;
```

REML not available in NLMIXED. For comparison purposes, I will use ML, and then repeat with REML

Decayed, Missing, Filled Teeth

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Estimate (ML)	Estimate (REML)
patient	41.8611	50.5528
Residual	6.3889	6.3889

Decayed, Missing, Filled Teeth

```
proc sql ;
   create table icc as
   select sum(estimate*(covparm='patient'))
          / sum(estimate) as icc
         from cov1 ;
quit ;
proc print data = icc ; run ;
```


```
Obs     icc (ML)     icc (REML)
1       0.86759        0.88780
```

Decayed, Missing, Filled Teeth

PROC NLMIXED Approach

```
PROC NLMIXED data=dental method=firo  ;
    parms mu=10  s_subj = 50 s_err = 6 ;
    pred = mu + beta ;
    model score ~ normal(pred, s_err) ;
    random beta ~ normal(0, s_subj)
            subject = patient ;
            estimate 'icc'
                     s_subj/(s_subj+s_err);
run ;
```

```
Label Estimate
icc   0.8676
```

Same estimate as PROC MIXED using ML

Don’t want to do the programming?

- There is a user-written macro that can be downloaded from the SAS website
- http://support.sas.com/kb/25/031.html
- The macro is called %INTRACC
- Computes 6 different versions of the ICC

%INTRACC and the Dental Data

- **Macro invocation:**

```
%intracc(depvar=score,
         target=patient,
         rater=rater, nrater=4)
```

- **Documentation states that macro computes 6 different versions of ICC**
- **Output includes 9 versions**

%INTRACC Output

Winer reliability: single score	Winer reliability: mean of k scores	Winer reliability mean of 4 scores	Shrout-Fleiss reliability: single score	Shrout-Fleiss reliability: random set
0.88780	0.96937	0.96937	0.88780	0.88958

Shrout-Fleiss reliability: fixed set	Shrout-Fleiss reliability: mean k scores	Shrout-Fleiss rel: rand set mean k scrs	Shrout-Fleiss rel: fxd set mean k scrs
0.94996	0.96937	0.96990	0.98700

- Same ICC value as PROC MIXED using REML

Dental Data Using PROC GLM

- First, output the mean squares to a dataset
- Next, examine expected mean squares to determine relationship between MS's and variance components, σ_β^2 and σ^2

GLM Code

```
ods output ModelANOVA=ms1
     (where=(hypothesistype=3)) ;
ods output OverallANOVA=ms0
     (where=(source='Error')) ;
proc glm data = dental  ;
      class patient ;
     model score = patient ;
     random patient ;
run ;
```

GLM Continued

```
The GLM Procedure
Source      Type III Expected Mean Square
patient     Var(Error) + 4 Var(patient)
```

σ^2 σ_β^2

- So, use MS(error) to estimate σ^2
- Use [MS(patients) – MS(error)]/4 to estimate σ_β^2

GLM Continued Again

```
proc sql ;
 create table icc as
 select a.ms as s_err,(b.ms-a.ms)/4 as s_sub,
        ((b.ms-a.ms)/4)/(a.ms+(b.ms-a.ms)/4)
        as icc
        from ms0 a, ms1 b;
quit ;
proc print data = icc ; run ;
```

Same estimates as PROC MIXED with REML

Obs	s_err	s_subj	icc
1	6.388889	50.5528	0.88780

Comparison for RCB Design

- GLM is cumbersome!
- %INTRACC is too much of a black box for me
- NLMIXED is simple, which I like, but no option to do REML
- MIXED seems to me like best compromise
- What about more complicated designs?

Random Rater Effect

- **Suppose same 4 raters assess each subject, but that the 4 raters randomly selected from larger pop'n of raters**

$$Y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

$\beta_i \sim N(0, \sigma_\beta^2)$ random subject effect

$\tau_j \sim N(0, \sigma_\tau^2)$ random rater effect

$\varepsilon_{ij} \sim N(0, \sigma^2)$ experimental error

ICC for Random Rater Effect

$$Cov(Y_{ij}, Y_{ik}) = E(Y_{ij}Y_{ik}) - E(Y_{ij})E(Y_{ik})$$
$$= \mu^2 + E(\beta_i^2) - \mu^2$$
$$= \sigma_\beta^2$$

$$Var(Y_{ij}) = Var(Y_{ij}) = \sigma_\beta^2 + \sigma_\tau^2 + \sigma^2$$

$$\therefore Corr(Y_{ij}, Y_{ik}) = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\tau^2 + \sigma^2} = ICC$$

NLMIXED and GLM

- Oops!! NLMIXED can't handle "crossed" random effects. Maybe in an upcoming version?
- Here are the expected mean squares from PROC GLM:

```
                The GLM Procedure
Source          Type III Expected Mean Square
patient           Var(Error) + 4 Var(patient)
rater             Var(Error) + 6 Var(rater)
```

I'm too lazy to do the calculations so I'll skip ahead to PROC MIXED

MIXED For Random Rater Effect

```
ods output CovParms = cov2 ;
proc mixed data = dental method = REML ;
      class patient rater ;
      model score =  ;
      random patient rater ;
run ;
proc sql ;
      create table icc2 as
      select sum(estimate*(covparm='patient'))
      / sum(estimate) as icc
      from cov2 ; quit ;
```

Obs	icc
1	0.88958

%INTRACC Output

Winer reliability: single score	Winer reliability: mean of k scores	Winer reliability: mean of 4 scores	Shrout-Fleiss reliability: single score	Shrout-Fleiss reliability: random set
0.88780	0.96937	0.96937	0.88780	0.88958

Shrout-Fleiss reliability: fixed set	Shrout-Fleiss reliability: mean k scores	Shrout-Fleiss rel: rand set mean k scrs	Shrout-Fleiss rel: fxd set mean k scrs
0.94996	0.96937	0.96990	0.98700

- Same macro call as earlier & same output
- Same value as MIXED with REML

Raters from 2 Different Specialties

- Suppose that 1st two raters are from one dental specialty & 2nd two raters are from another
- Only two specialities of interest
- ∴ speciality is a fixed effect
- 2 raters picked at random from each specialty
- Rater effect expected to differ by specialty

Model for Different Specialties

$$Y_{ijk} = \mu + \beta_i + \tau_{j(k)} + \gamma_k + \varepsilon_{ij(k)}$$

$\beta_i \sim N(0, \sigma_\beta^2)$ **random subject effect**

$\tau_{j(1)} \sim N(0, \sigma_{\tau 1}^2)$ **random rater effect for 1st specialty**

$\tau_{j(2)} \sim N(0, \sigma_{\tau 2}^2)$ **random rater effect for 2nd specialty**

γ_k **fixed specialty effect**

$\varepsilon_{ij} \sim N(0, \sigma^2)$ **experimental error**

ICC for this Model

Same subject, same specialty:

$$ICC = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_{\tau k}^2 + \sigma^2}$$

Same subject, different specialties:

$$ICC = \frac{\sigma_\beta^2}{\sqrt{\sigma_\beta^2 + \sigma_{\tau 1}^2 + \sigma^2} \times \sqrt{\sigma_\beta^2 + \sigma_{\tau 2}^2 + \sigma^2}}$$

Challenge

- Expressions for Model and ICC are easy to write out, but how to fit the model and estimate ICC?
- "group=specialty" option in MIXED varies *all* parameters by specialty instead of just the rater variance
- NLMIXED would allow this if there was only one random effect, but there are 2, which it can't handle