EVALUATION AND COMPARISON OF METHODS OF MEASUREMENTS

# DAY 2

## Advanced comparison of methods of measurements

Niels Trolle Andersen and Mogens Erlandsen

mogens@biostat.au.dk

Department of Biostatistics



## DAY 2

**xtmixed**: a short introduction

**Evaluation of a method with repeated measurements**

Example 2: 5 repeated measurements performed 4 times

- Training effect
- Variance components (different sources of errors)

**Comparison of more than 2 methods**

Example 3: 3 methods, one considered as gold standard

- Correlated measurement errors
- Comparison of standard deviation on different methods



## xtmixed: a short introduction

Example 1 (DAY 1), compare two measurements of Achilles tendon thickness by observer **a**.

Correlation between the two measurements:
**pwcorr a1 a2** => corr(a1,a2) = 0.8738

That is: Data in "long" format cannot be independent?

In **xtmixed** the NON-independent behaviour is created by a **variance component,** i.e. a random component **shared** among the two measurements that should be dependent.



## xtmixed: a short introduction

Model 3: $y_{i,j} = \mu + \delta_j + P_i + E_{i,j} \qquad j = 1,2 \qquad i = 1,...,46$

$\mu$: Overall (average) level

$\delta_j$: Systematic difference between measurement 1 and 2

$P_i$: Variance component for each individual

$E_{i,j}$: Measurement error (individual x measurement)

Fixed effects: $\mu, \delta_j$

Random effects: $P_i, E_{i,j}$ $\Rightarrow$ **MIXED** effect model

Assumptions: $P_i \sim N(0, \sigma_P^2)$, $E_{i,j} \sim N(0, \sigma_E^2)$ — All random components are **INDEPENDENT!**

## **xtmixed**: a short introduction

In this model each pair of observations, i.e. $(y_{i,1}, y_{i,2})$ are correlated, and

$$\text{corr}(y_{i,1}, y_{i,2}) = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_E^2} = 0.87 \approx$$ Pearson's ***r***

Stata output:

| Random-effects Parameters | Estimate | Std. Err. | [95% Conf. | Interval] |
|---|---|---|---|---|
| id: Identity | | | | |
| var(_cons) | .5307923 | .1205127 | .3401459 | .8282929 |
| var(Residual) | .0789686 | .016648 | .0522403 | .1193723 |



## EXAMPLE 2

DATA:

40 healthy persons were tested for "one-leg static balance". The outcome of the test is a "balance index" (arbitrary units), we will look at the natural logarithm to the index (ln_bi).

Each person tested and retested 1 month later

On each test day the test person completed 2 test series (with 30 minutes interval between the 2 series).

Each test series: 5 measurements which lasted 20 seconds each (1 minute pause between measurements).

Total number of observations: N = 40 x 2 x 2 x 5 = 800



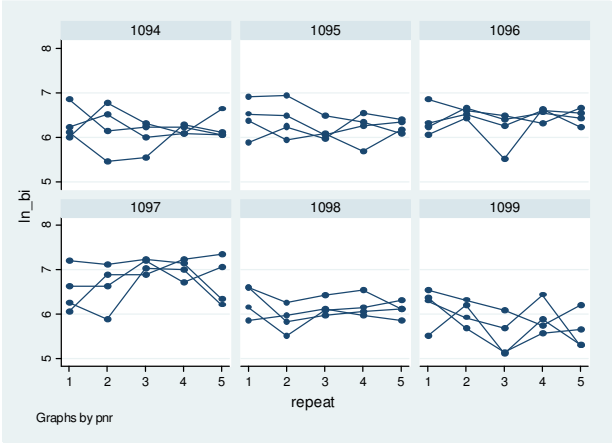
## EXAMPLE 2

Extract of data: Data format "long"

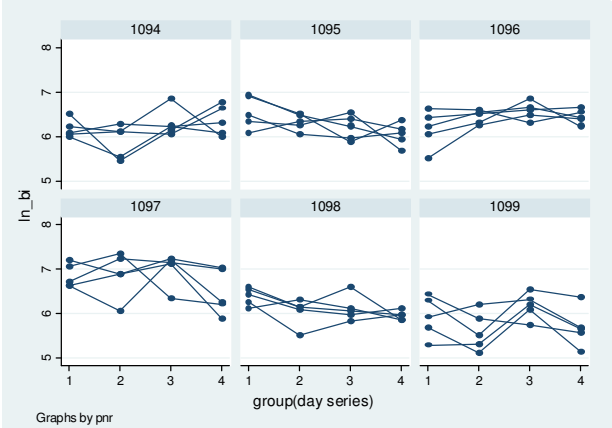
| pnr | repeat | series | day | ln_bi |
|---|---|---|---|---|
| … | | | | |
| 1094 | 5 | 1 | 2 | 6.052089 |
| 1094 | 1 | 2 | 2 | 5.998937 |
| 1094 | 2 | 2 | 2 | 6.747587 |
| 1094 | 3 | 2 | 2 | 6.298949 |
| 1094 | 4 | 2 | 2 | 6.068426 |
| 1094 | 5 | 2 | 2 | 6.624065 |
| 1095 | 1 | 1 | 1 | 6.897705 |
| 1095 | 2 | 1 | 1 | 6.942157 |
| 1095 | 3 | 1 | 1 | 6.46925 |
| 1095 | 4 | 1 | 1 | 6.342122 |
| 1095 | 5 | 1 | 1 | 6.082219 |
| … | | | | |



## EXAMPLE 2: Plot of repeat within series

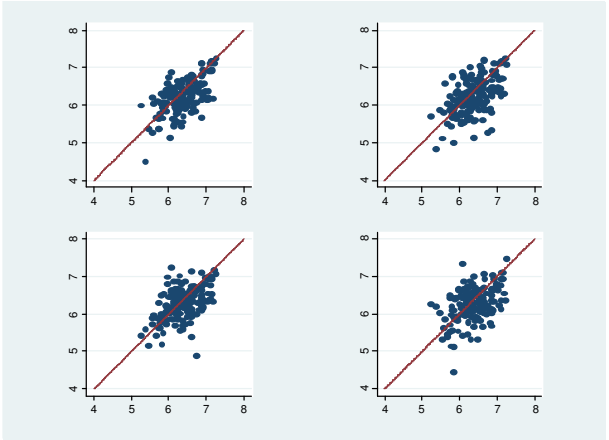


## EXAMPLE 2: Plot of series for each repeat



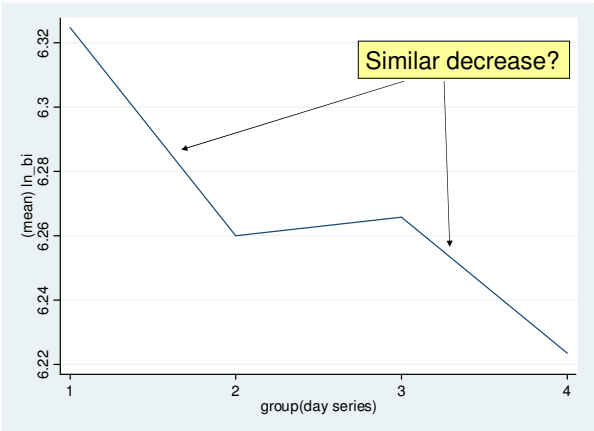



## EXAMPLE 2: Plot of repeat 2, 3, 4, 5 vs 1



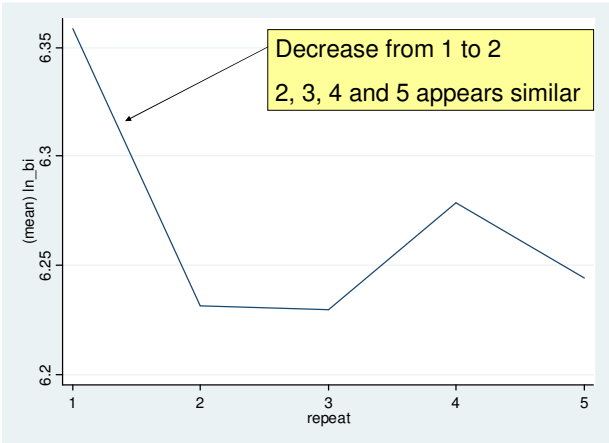


## EXAMPLE 2: Plot of average pr. session






## EXAMPLE 2: Plot of average pr. repeat



## EXAMPLE 2: Random effects

Repeated observations in several ways, - observations are not independent…

Classical analysis of variance can handle designs of this type (balanced designs). But this approach has at least two drawbacks: missing data and/or unbalanced designs.

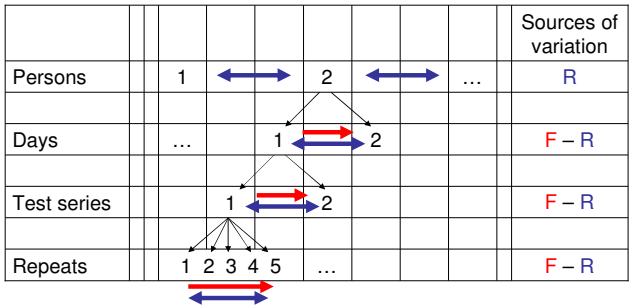
We need another way to specify more than one source of random variation.

Mixed models (Stata: **`xtmixed`**) can handle both these drawbacks, - but not always easy to set up in the proper way.



## EXAMPLE 2

Static balance test: each person measured on two different days, 2 test series, 5 repeated meas.




Fixed (systematic) effect ="training" effect ?

Random effect



## EXAMPLE 2: Mixed models

Fixed effects: effects common to groups of observations

Random effects: random deviations of the observations from the sum of fixed effects (predicted values)

In symbolic language:

Observation = Fixed effects + [Random effects]

= Day + Series + Repeat (we assume no interactions!)

+ [Pnr] + [Pnr x Day] + [Pnr x Day x Series]

+ [Pnr x Day x Series x Repeat]

A **mixed model** includes both **fixed effects** and **random effects**



## EXAMPLE 2: Random effects

Random effects - interpretation:

[Pnr]: Inter-individual variation

[Pnr x Day]: Intra-individual variation between days (longterm random variation)

[Pnr x Day x Series]:

Intraindividual variation between series within day (shortterm random variation)

[Pnr x Day x Series x Repeat]:

Intraindividual variation within series = i.e among repeats within a series = Residual

The combination of Pnr, Day, Series and Repeat **uniquely** identify each single observation

## EXAMPLE 2: Stata: `xtmixed`

Mixed model:

**3 random effects** (alsop called variance components) in a hierarcical (nested) structure and **residual** variation

Stata:

```
… || pnr: || day: || series:
```

By default Stata includes the residual variation as the ”remaining variation”.

The full Stata command:

```
xi: xtmixed ln_bi ///
      i.day i.series i.repeat ///
      || pnr: || day: || series:
```



## EXAMPLE 2: `xtmixed/numbers`

Extract from output:

```
Mixed-effects REML regression          Number of obs      =       800

------------------------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------------------
            pnr |       40         20       20.0         20
            day |       80         10       10.0         10
         series |      160          5        5.0          5
------------------------------------------------------------------------------
```

Check this part carefully!

- Total number of observations = 800
- 2 days, 2 test series per day, 5 repeats … = 2 x 2 x 5 = 20
- 2 test series per day, 5 repeats per series = 2 x 5 = 10
- Number of repeats = 5



## EXAMPLE 2: `xtmixed/random effects`

Extract from output:

Dont use!

```
------------------------------------------------------------------------------
Random-effects
Parameters          |   Estimate   Std. Err.      [95% Conf. Interval]
--------------------+---------------------------------------------------------
pnr: Identity       |
          sd(_cons) |   .2761153   .0375038        .21158    .3603348
--------------------+---------------------------------------------------------
day: Identity       |
          sd(_cons) |   .1225056   .0274793      .0789264     .190147
--------------------+---------------------------------------------------------
series: Identity    |
          sd(_cons) |   .0862984    .025167      .0487269      .15284
--------------------+---------------------------------------------------------
       sd(Residual) |   .3083595    .008646       .291871    .3257796
------------------------------------------------------------------------------
```

The estimated variance components, - expressed as standard deviations



## EXAMPLE 2 - interpretation

Variance components play an important role in design of future experiments.

Example: Design an experiment to compare difference in Balance Index after two ”treatments”. Treatments given at two different days. Use a paired experiment (cross-over). Response = daily average of ln_bi.

| Design (per treatment) | Days | Series | Repeats |
|---|---|---|---|
| S(imple) | 1 | 1 | 1 |
| A | 1 | 1 | 5 |
| B | 1 | 2 | 3 |

## EXAMPLE 2 – interpretation/designs

Remember: standard deviations can't be added, but variances can. [option **`var`** generates variances]

$$\begin{aligned}\text{Var("S-dif")} &= 2\times(\text{Var(day)}+\text{Var(series)}+\text{Var(Residual)})\\ &= 2\times(\text{sd}^2\text{(day)}+\text{sd}^2\text{(series)}+\text{sd}^2\text{(Residual)})\\ &= 2\times(0.123^2+0.086^2+0.308^2)\\ &= 2\times 0.1175 = 0.2350\end{aligned}$$

$$\text{sd("S-dif")} = \sqrt{0.2350} = 0.485$$

$$\begin{aligned}\text{Var("A-dif")} &= 2\times(\text{Var(day)}+\text{Var(series)}+\text{Var(Residual)/5})\\ &= 2\times(\text{sd}^2\text{(day)}+\text{sd}^2\text{(series)}+\text{sd}^2\text{(Residual)/5})\\ &= 2\times(0.123^2+0.086^2+0.308^2/5)\\ &= 2\times 0.0414 = 0.0828\end{aligned}$$

$$\text{sd("A-dif")} = 0.288$$



## EXAMPLE 2 – interpretation/designs

$$\begin{aligned}\text{Var("B-dif")} &= 2\times\left(\text{Var(day)}+(\text{Var(series)}+\text{Var(Residual)/3})/2\right)\\ &= 2\times\left(\text{sd}^2\text{(day)}+(\text{sd}^2\text{(series)}+\text{sd}^2\text{(Residual)}/3)/2\right)\\ &= 2\times\left(0.123^2+(0.086^2+0.308^2/3)/2\right)\\ &= 2\times 0.0346 = 0.0692\end{aligned}$$

$\text{sd("B-dif")} = 0.263$ (Not much gained by repeated series)

In design considerations:

- Order variance components by magnitude (here repeat > pnr > day > series)
- Make repeated observations on the largest components (here pnr and repeat)



## EXAMPLE 2 - interpretation

Total variance (only one measurement):

$$\begin{aligned}\text{Var(Total)} &= \text{Var(pnr)}+\text{Var(day)}+\text{Var(series)}+\text{Var(Residual)}\\ &= \text{sd}^2\text{(pnr)}+\text{sd}^2\text{(day)}+\text{sd}^2\text{(series)}+\text{sd}^2\text{(Residual)}\\ &= 0.276^2+0.123^2+0.086^2+0.308^2\\ &= 0.1938\end{aligned}$$

Hence: $\text{sd(Total)} = \sqrt{0.1938} = 0.440$

Select one obs per person

```
. sum ln_bi if day==1 & series==1 & repeat==3
Variable |   Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
   ln_bi |    40    6.263047    .4403927   5.501258   7.194437
```

Variance components are sometimes expressed as percentage of total variance (compare with ICC, Day 1).



## EXAMPLE 2: `xtmixed/fixed effects (1)`

Extract from output (1):

Compare with p. 11

```
------------------------------------------------------------------------------
      ln_bi |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
    _Iday_2 |  -.0476178   .0375765    -1.27   0.205    -.1212664     .0260309
_Iseries_2 |  -.0536296   .0257219    -2.08   0.037    -.1040435    -.0032157
------------------------------------------------------------------------------
```

Conclusion (1):

No difference between day 1 and 2 (p=0.2)
(Small?) difference between series 1 and 2 (p=0.037).

Note:
`Std. Err(_Iday_2) > Std. Err.(_Iseries_2)`
Why?

## EXAMPLE 2: `xtmixed/fixed effects (2)`

Extract from output (2):

Compare with p. 12

| ln_bi | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _Irepeat_2 | -.1268038 | .0344756 | -3.68 | 0.000 | -.1943748 | -.0592328 |
| _Irepeat_3 | -.1287257 | .0344756 | -3.73 | 0.000 | -.1962967 | -.0611546 |
| _Irepeat_4 | -.0800571 | .0344756 | -2.32 | 0.020 | -.1476281 | -.0124861 |
| _Irepeat_5 | -.1143186 | .0344756 | -3.32 | 0.001 | -.1818896 | -.0467476 |
| _cons | 6.409089 | .0570335 | 112.37 | 0.000 | 6.297306 | 6.520873 |

Conclusion (2):

Repeat 1 (reference) appears to be different.
(Comparison of repeat 2 to 5 gives p=0.46)

Could consider to exclude repeat 1?

But that is not (solely) the decision of the statistician!



## EXAMPLE 2 – Checking assumptions

Checking assumptions:

Residuals can be obtained and checked as usual.

Other random effects (pnr, day, series) can be obtained as socalled BLUP's (**B**est **L**inear **U**nbiased **P**redictor), and checked with probability plots

Predicted values (including both fixed effects and BLUP's) can be obtained and checked as usual. Note: Residuals correspond to these predicted values, i.e.

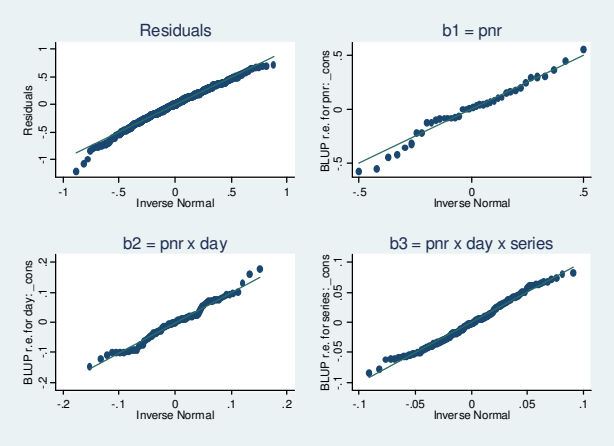
Observation = Prediction + Residual

Stata – postestimation:

```
predict res, residuals
predict xb, xb
predict fitted, fitted
predict b*, reffects
```

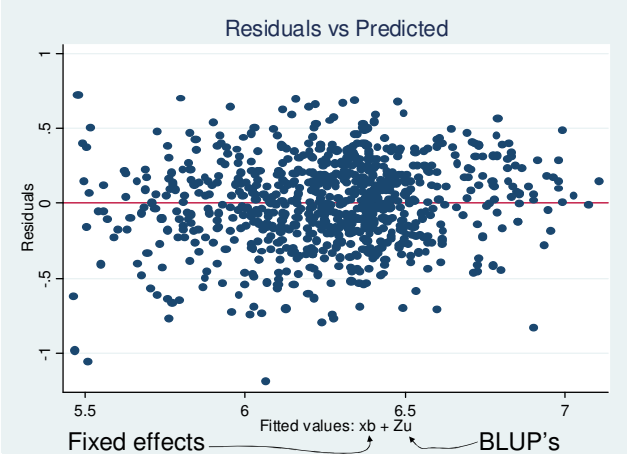


## EXAMPLE 2 – Checking assumptions






## EXAMPLE 2 – Checking assumptions



## EXAMPLE 2 – Checking assumptions

Could the random variation change from Day 1 to Day 2?

Make a separate analysis for each and look at the estimated standard deviations. Note: in this model the variable **day** should be removed from the **xtmixed** command.

```
------------------------------------------------------------------------------
Random-effects
Parameters          | Day 1 + 2           Day 1           Day 2
--------------------+---------------------------------------------------------
pnr: Identity       |
         sd(_cons) |   .2761             0.3067          0.2969
--------------------+---------------------------------------------------------
day: Identity       |
         sd(_cons) |   .1225             N/A             N/A
--------------------+---------------------------------------------------------
series: Identity    |
         sd(_cons) |   .0862             0.0814          0.0957
--------------------+---------------------------------------------------------
       sd(Residual) |   .3084             0.3060          0.3079
------------------------------------------------------------------------------
```



## XTMIXED/FIXED EFFECTS

Fixed effects can be specified ”as usual” in regression analysis. Categorical variables are entered using the Stata ”indicator”-syntax

**xi: xtmixed … i.cat1 i.cat2 …**

Interaction terms are allowed, i.e. …**i.cat1*i.cat2**

Continuous covariates may also be entered.

Fixed effects: Stata computes Wald type test, not optimal with sample small sizes (p-values will be too small). In this example only minor difference (e.g. p=0.037 would be 0.04).



## XTMIXED/RANDOM EFFECTS

What can goes wrong?

```
Mixed-effects REML regression           Number of obs     =        800

------------------------------------------------------------------------------
                |   No. of       Observations per Group
 Group Variable |   Groups    Minimum    Average    Maximum
----------------+-------------------------------------------------------------
            pnr |       40         20       20.0         20
            day |       80         10       10.0         10
         series |      160          5        5.0          5
------------------------------------------------------------------------------
```

Number of BLUP’s for the 3 variance components: (b1, b2, b3)

Should be a ”reasonable” number in each group ( > 5/10)

Otherwise, consider the effect as fixed



## XTMIXED/RANDOM EFFECTS

What can goes wrong?

Random effects are **RANDOM**, i.e.

Random effects represent something random:

- Random sample of individuals/patient
- Randomly chosen days (but fixed interval between)
- Randomly chosen series within days

**If not**, the effects should be **fixed**

How to do:
1. Make the fixed effects part work without random effects
2. Add 1. random component (pnr), make it work
3. Add 2. random component (day), make it work …

## EXAMPLE 3

Measurement with 3 different methods

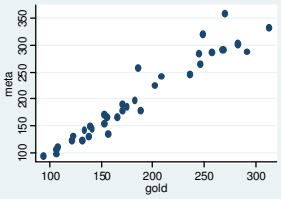
- Two methods: MetA and MetB are similar
- One method (Gold) is considered as 'gold standard'
- N = 34, each person measured with all 3 methods

Extract of data:

| gold | meta | metb | ptnr |
|---|---|---|---|
| 93.62815 | 95.08382 | 103.2247 | 1 |
| 106.1439 | 105.136 | 113.5475 | 2 |
| 106.5172 | 96.65213 | 103.328 | 3 |
| 107.8929 | 111.5212 | 113.6232 | 4 |
| . . . . | | | |



## EXAMPLE 3: scatterplots

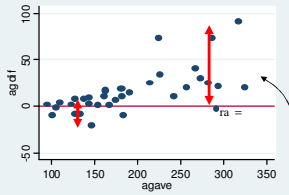



3 bivariate scatter plots

- Methods clearly associated



## EXAMPLE 3: dif vs. ave plots



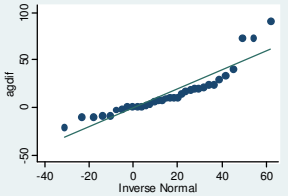

Clear trend

Random variation on **dif** depend on **ave**

```
agdif = meta - gold
bgdif = metb - gold
badif = metb - meta
ave(-rage) defined similarly
```

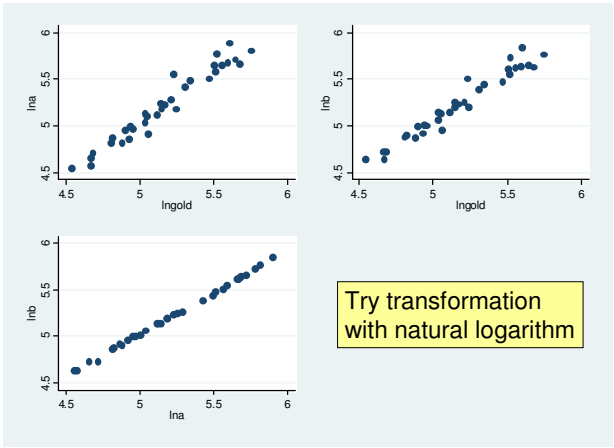


## EXAMPLE 3: dif – prob.plots




We can do better!

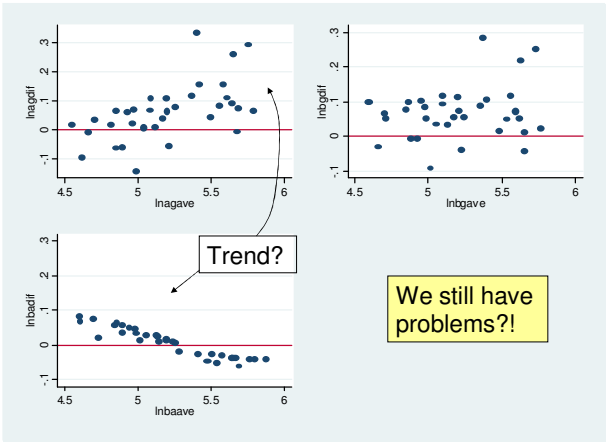
## EXAMPLE 3: scatterplots on ln-values




Try transformation with natural logarithm



## EXAMPLE 3: dif vs. ave plots on ln-values

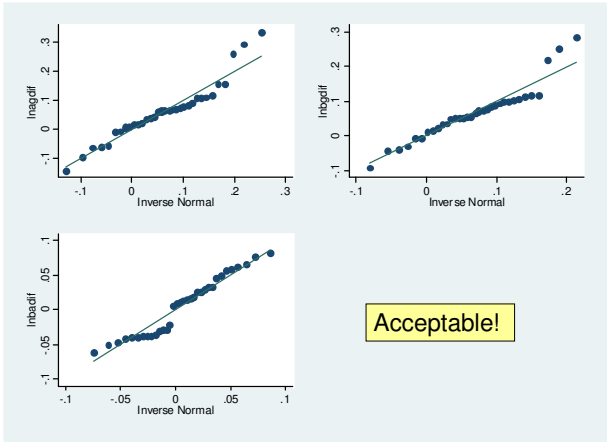



Trend?

We still have problems?!



## EXAMPLE 3: dif – prob.plots on ln-values




Acceptable!



## EXAMPLE 3: ttest on dif – ln-values

Stata: `ttest "dif"` (reduced output)

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnagdif | 34 | .060879 | .0171647 | .1000865 | .0259572 | .0958008 |
| lnbgdif | 34 | .0665448 | .0132726 | .0773922 | .0395414 | .0935482 |
| lnbadif | 34 | .0056658 | .0072489 | .0422682 | -.0090823 | .0204139 |

Pr(|T| > |t|) = 0.0012 (lnagdif) 0.0000 (lnbgdif) 0.4400 (lnbadif)

Paired t-test:

Systematic difference between both MetA and MetB and the gold standard. Calibration problem???

No systematic difference between MetA and MetB.

## EXAMPLE 3: ttest on dif – ln-values

Stata: `ttest "dif"` (reduced output)

| Variable | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| lnagdif | 34 | .060879 | .0171647 | .1000865 | .0259572 | .0958008 |
| lnbgdif | 34 | .0665448 | .0132726 | .0773922 | .0395414 | .0935482 |
| lnbadif | 34 | .0056658 | .0072489 | .0422682 | -.0090823 | .0204139 |

Pr(|T| > |t|) = 0.0012 (lnagdif) 0.0000 (lnbgdif) 0.4400 (lnbadif)

The standard deviations on the differences (i.e. the random variation) appear to be different?

Cannot compare the standard deviations by Stata (e.g. `sdtest`), - the differences are not statistically independent (derived from the same data).



## EXAMPLE 3: standard deviations

This design cannot estimate the standard deviation on the measurement error for each method.

Can only estimate the standard deviation on the pairwise difference between two of the methods, i.e.

sd(lnagdif), sd(lnbgdif) and sd(lnbadif)

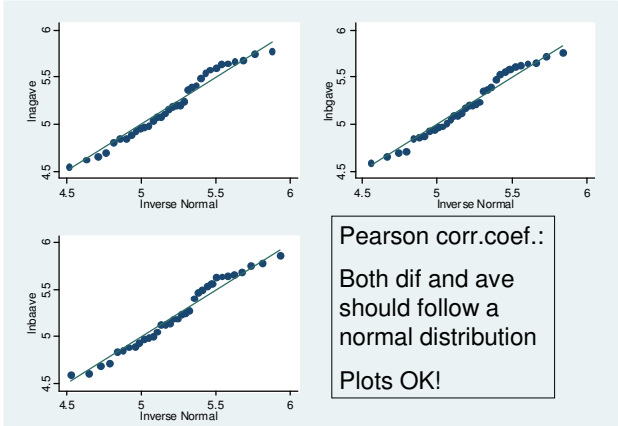
What happens if the standard deviations on the measurement errors differ between the 3 methods?

The scatter plot of dif vs. ave will show a **trend**!
i.e. **dif and ave will be correlated!**

Thus, check the Pearson correlation coefficient!



## EXAMPLE 3: ave – prob.plots on ln-values




Pearson corr.coef.:

Both dif and ave should follow a normal distribution

Plots OK!



## EXAMPLE 3: Pearson's corr.coef

`pwcorr lnagdif lnagave, sig`

| | lnagdif | lnagave |
|---|---|---|
| lnagdif | 1.0000 | |
| lnagave | 0.5729 | 1.0000 |
| | 0.0004 | |

MetA vs Gold
Correlation coefficient

p-value (H: corr.coef=0)

We reject!

`pwcorr lnbgdif lnbgave, sig`

| | lnbgdif | lnbgave |
|---|---|---|
| lnbgdif | 1.0000 | |
| lnbgave | 0.2237 | 1.0000 |
| | 0.2035 | |

MetB vs Gold
Correlation coefficient

p-value (H: corr.coef=0)

We accept!

## EXAMPLE 3: MetA vs. MetB

```
pwcorr lnbadif lnbaave, sig
```

```
             | lnbadif   lnbaave
-------------+------------------
   lnbadif   |   1.0000
             |
             |
   lnbaave   |  -0.9408   1.0000
             |   0.0000
```

MetB vs MetA
Correlation coefficient

p-value (H: corr.coef=0)
We reject!

Remember:

sd(lnagdif) = 0.100

sd(lnbgdif) = 0.077

sd(lnbadif) = **0.042**

Paired t-test (A vs B) = 0.22

Apparently, MetA and MetB do agree better than with the gold standard?!

- They may be equally bad!
- Look at the dif-ave plot (A vs B), p. 38



## EXAMPLE 3: A-G vs. B-G

Comparison of Method A and B could also be done as a comparison of their deviations from the gold standard using the Bland-Altman analysis from DAY 1:

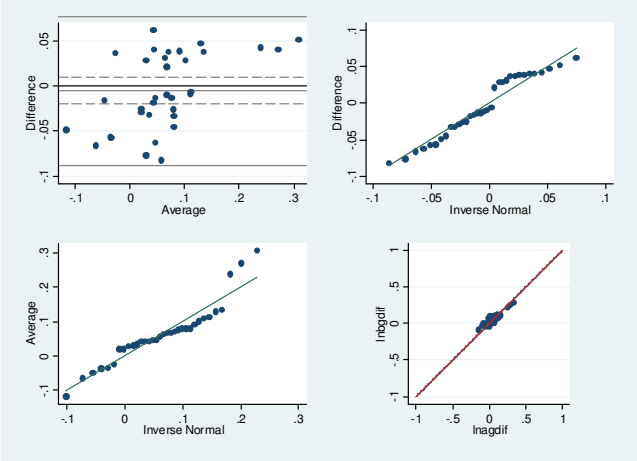
```
BAanalysis lnagdif lnbgdif, diag
```

Reduced output:

```
Mean difference (bias lnagdif - lnbgdif): -0.006
                                          (CI -0.020 to  0.009  , p = 0.440)
Estimated sd on differences:  0.042 (CI  0.034 to  0.056)

Correlation between difference and average: Pearsons r =  0.548, p = 0.001
Correlation between difference and average: Spearmans rho =  0.564, p = 0.001

.....

Comparison of two methods:
Estimated with in subject sd:  0.084 (variance  0.007)
Estimated additional error sd (lnagdif):  0.054 (variance  0.003)
Estimated additional error sd (lnbgdif):      . (variance -0.001)
Low or negative error variance indicates that model assumptions are violated.
```



## EXAMPLE 3: A-G vs. B-G






## EXAMPLE 3: Conclusion

**Conclusion:**

The standard deviation on the measurement error generated by method A appears to be different from the gold standard (p=0.0004), whereas method B has the same standard deviation (p=0.20).

The correlation between the two set of differences lnagdif and lnbgdif is statistically different from 0. Since sd(lnagdif)=0.100 and sd(lnbgdif)=0.077, Method A has greater measurement error than method B.

The measurements errors by methods A and B appears to be correlated.