Randomized controlled trials with unobserved heterogeneity among subjects

Rasmus Waagepetersen

October 14, 2025

Outline:

1. Baseline adjustment
2. Linear mixed model
3. Misspecified linear model

RCT with baseline measurement

Consider RCT with observations Y_{aij}, $i = 1, \ldots, n_a$, $j = b, e$ $i = 1, \ldots, n_a$, for $a = T$ (the treatment arm) or $a = C$ (the control arm).

The observations Y_{aib} are baseline measurements of the outcome variable before treatment or placebo is administered. The observations Y_{aie} are the endpoint values recorded after treatment or placebo is given.

We assume the pairs (Y_{aib}, Y_{aie}), $a = T, C$, $i = 1, \ldots, n_a$ (one for each subject) are independent but Y_{aib} and Y_{aie} could be dependent.

Following the logic of an RCT, we let P_b denote the common distribution of all baseline measurements and P_a denote the common distribution of Y_{aie}, $a = T, C$.

Letting Y_b, Y_T and Y_C denote generic random variables following these three distributions, our target of estimation is the average treatment effect

$$\psi = \mathbb{E}Y_T - \mathbb{E}Y_C$$

Recall from Emilie's lectures that this is an estimate of a causal treatment effect due to randomization.

Strategies for estimation of ψ

The obvious estimate of ψ is the simple difference of averages

$$\hat{\psi} = \bar{Y}_{T \cdot e} - \bar{Y}_{C \cdot e}.$$

The question is whether we can do better by using the baseline values ?

The answer is yes if the baseline and endpoint measurement for a subject are correlated.

Mixed model

We consider a simple mixed model as framework for our discussion:

$$
\begin{aligned}
Y_{aib} &= \mu + U_{ai} + \epsilon_{aib} \\
Y_{aie} &= \mu + \gamma + \psi 1[a = T] + U_{ai} + \epsilon_{aie}
\end{aligned} \tag{1}
$$

where the U_{ai} and the ϵ_{aij} are independent with variances τ^2 for the U_{ai} and σ^2 for the ϵ_{aij}.

The random effects U_{ai} model random heterogeneity *between* subjects.

γ is a "time" effect and we could add further covariate effects but skip for simplicity.

Differencing/"paired t-test"

Define $\Delta_{ai} = Y_{aie} - Y_{aib}$. Then for the preceding example,

$$\mathbb{E}\Delta_{ai} = \psi 1[a = T] \quad \mathbb{V}\mathrm{ar}\Delta_{ai} = 2\sigma^2$$

We may estimate ψ by

$$\hat{\psi}_d = \bar{\Delta}_{T\cdot} - \bar{\Delta}_{C\cdot}$$

Assuming mixed model is valid, when is this advantageous compared to simple difference ?

(considering possible values of τ^2 and σ^2) ?

Adjusting for baseline in a linear regression model

Obtain least squares estimate $\hat{\psi}_l$ based on linear regression model with baseline as covariate:

$$Y_{aie} = \alpha_0 + \alpha_1 Y_{aib} + \psi 1[a = T] + \varepsilon_{aie} \tag{2}$$

According to Schuler's book, in case of an RCT, this gives a consistent estimator for ψ regardless of whether linear regression is misspecified or not (we will return to this later)

Linear mixed model estimate

Assume known τ^2 and σ^2.

Under the linear mixed model, the maximum likelihood estimate of $\beta = (\mu, \gamma, \psi)$ is the weighted least squares estimate

$$\hat{\beta} = (\mathsf{X}^\mathsf{T}\Sigma^{-1}\mathsf{X})^\mathsf{T}\mathsf{X}^\mathsf{T}\Sigma^{-1}\mathsf{Y}$$

where Σ is the (2×2 block diagonal) covariance matrix of the observation vector $\mathsf{Y} = (Y_{C1b}, Y_{C1e}, \ldots, Y_{Cn_Cb}, Y_{Cn_Ce}, Y_{T1b}, Y_{T1e}, \ldots, Y_{Tn_Tb}, Y_{Tn_Te})^\mathsf{T}$, and X is the design matrix with columns $1_{2n_C+2n_T}$, $(0, 1, 0, \ldots, 0, 1)$ and $(0, \ldots, 0, 0, 1, \ldots, 0, 1)^T$.

One can show (exercise) that the WLS of $\psi =$ is an unbiased estimate of the treatment effect and with same variance as for the linear model estimate adjusting for baseline.

Case study: fraction test results for 4th grade students

Computing with fractions is a key obstacle for primary school students

Consider randomized trial to test a new math teaching system against current practice.

125 schools were randomly allocated to current practice or new teaching system. 6589 students participate in the trial.

Consider data Autumn 22 (baseline) and Spring 23 (one school year of treatment).

Fixed effects model

Parameters: main effect of teaching method (ψ) and main effect of participating in school year (γ)

Expected values in two-way table:

	A22 (baseline)	S23
Current	μ	$\mu + \gamma$
New	μ	$\mu + \gamma + \psi$

Least squares analysis for ordinary two-way ANOVA

```
#create custom variable for treatment effect
> Treateffect=as.numeric(scores$Test==2 & scores$treatmentlabel=="NEW")
> fitbroklm=lm(Broker~Testlbl+Treateffect,data=scores,na.action=na.excl
> summary(fitbroklm)

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.7092     0.1995 123.841   <2e-16 ***
TestlblF2023  11.3659     0.3491  32.559   <2e-16 ***
Treateffect   -0.2957     0.4070  -0.727    0.468
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.33 on 10116 degrees of freedom

> 14.33^2 # residual variance
[1] 205.3489
```

Residual variance 205.3

Conclusions ? Any potential problems with this analysis ?

Simple estimate (difference of means)

```
> Treat=Broker[scores$Test==2 & scores$treatment==1]
> Control=Broker[scores$Test==2 & scores$treatment==0]
> t.test(Treat,Control)

Welch Two Sample t-test

data:  Treat and Control
t = -0.67237, df = 4957.7, p-value = 0.5014
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.1578986  0.5664905
sample estimates:
mean of x mean of y
 35.77941  36.07511

> mean(na.omit(Treat))-mean(na.omit(Control))
[1] -0.2957041
```

Result similar to previous analysis

Analysis based on differences

Similar to paired t-test we may consider the difference between spring and autumn test for each student.

This eliminates parameter μ and possible random effects !

Difference for student with current (assuming linear mixed model):

$$\begin{aligned} Y_{i2} - Y_{i1} =& \mu + \gamma + U_i + U_{class(i)} + U_{school(i)} + \epsilon_{i2} \\ & -(\mu + U_i + U_{class(i)} + U_{school(i)} + \epsilon_{i1}) = \gamma + \tilde{\epsilon}_i \end{aligned}$$

Similarly, for student with new:

$$Y_{i2} - Y_{i1} = \gamma + \psi + \tilde{\epsilon}_i$$

Least squares analysis based on differences

```
> fitdiffbroklm=lm(diffbrok~treatment.x,data=scoresdiff)
> summary(fitdiffbroklm)

Call:
lm(formula = diffbrok ~ treatment.x, data = scoresdiff)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.6640     0.2247  47.464  < 2e-16 *** #a
treatment.x   1.6723     0.3191   5.241 1.67e-07 *** #g
```

Estimated treatment effect: 1.6723, highly significant.

Note: number of difference observations 4498 is less than half of previous number of observations.

Baseline adjusting

```
> fitbasebrok=lm(Broker.y~Broker.x+treatment.x,data=scoresdiff)
> summary(fitbasebrok)

Call:
lm(formula = Broker.y ~ Broker.x + treatment.x, data = scoresdiff)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.17705    0.38588  36.739  < 2e-16 ***
Broker.x     0.86525    0.01212  71.418  < 2e-16 ***
treatment.x  1.38508    0.31588   4.385 1.19e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.56 on 4495 degrees of freedom
  (1630 observations deleted due to missingness)
Multiple R-squared:  0.5317,Adjusted R-squared:  0.5314
F-statistic:  2551 on 2 and 4495 DF,  p-value: < 2.2e-16
```

Estimate now 1.38508

Again many observations left out.

Mixed model analysis

```
> fitbrok=lmer(Broker~Testlbl+Treateffect+(1|instnr/Klasse)+(1|ID),data=scores,na.action=na.exclude)
> summary(fitbrok)

Random effects:
 Groups        Name        Variance Std.Dev.
 ID            (Intercept) 126.401  11.243
 Klasse:instnr (Intercept)   5.488   2.343
 instnr        (Intercept)  17.329   4.163
 Residual                   57.058   7.554
Number of obs: 10119, groups:  ID, 5621; Klasse:instnr, 272; instnr, 119

Fixed effects:
             Estimate Std. Error        df t value Pr(>|t|)
(Intercept)   24.1797     0.4612  122.0087  52.425  < 2e-16 ***#mu
TestlblF2023  10.5816     0.2196 4845.3517  48.193  < 2e-16 ***#b
Treateffect    1.7166     0.3123 4866.3694   5.497 4.06e-08 ***#a

> 126.401+5.488+17.329+57.058#total variance
[1] 206.276
> (126.401+5.488+17.329)/206.276#Intra student correlation
[1] 0.72339 #correlation for two observations for same student
```

Same total variance as in two-way ANOVA (NB: does not make sense to add standard deviations)

Large correlation 0.72 between two test for same student

Positive significant effect of intervention. Result similar to estimate for pairwise difference approach.

Issue with pairwise difference and baseline adjustment approaches: loss of observations in cases where one test is missing for a student.

Not so easy to generalize if more than two tests (if several differences they will be correlated).

Model assumptions

Which approach is relying most on model assumptions ? (so far we have not assumed distributional properties in terms of normality or other specific distributions)

Or put in other words: which approach works under the weakest assumption?

A two-dimensional normally distributed vector (X, Y) can be factorized into regression of $Y|X$ and one-dimensional distribution of X.

So one might expect similar results for mixed model and baseline adjustment - at least if not too many missing observations...

The p-values produced by R procedures `lm` and `lmer` are based on assumption of normality.

However, least squares and weighted least squares estimates are unbiased regardless of assumption of normality

Model robust p-values can be obtained using sandwich estimator for standard errors (see Emilie's tutorial) or permutation tests.

Mixed model with repeated measurements

In practice a so-called mixed model for repeated measurements (MMRM) is often used for analysing clinical trials.

This is essentially just a linear normal model with a general covariance matrix for the residuals. That is,

$$(Y_{Cib}, Y_{Cie}) \sim N((\mu, \mu+\gamma)^\mathsf{T}, \Sigma) \quad (Y_{Tib}, Y_{Tie}) \sim N((\mu, \mu+\gamma+\psi)^\mathsf{T}, \Sigma)$$

where Σ is a general positive definite matrix. The simple linear mixed model is a special case of this.

Again further covariates could be added.

Decreasing variance by conditioning (Rao-Blackwellization)

We want to estimate $\psi = \mathbb{E}[Y_e|A=1] - \mathbb{E}[Y_e|A=0]$ which can be done by estimating $\mathbb{E}[Y_e|A=a]$, $a=0,1$.

Let W represent a vector of covariates that typically includes the baseline measurement Y_b. Then by the law of total expectation and by randomization,

$$\mathbb{E}[Y_e|A=a] = \mathbb{E}[\mathbb{E}[Y_e|A=a,W]|A=a] = \mathbb{E}[\mathbb{E}[Y_e|A=a,W]]$$

Suppose we have an unbiased estimate $\hat{\mathbb{E}}[Y_e|A=a,W]$ of the conditional expectation. Then we have the (plug-in) estimate

$$\hat{\mathbb{E}}[Y_e|A=a] = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbb{E}}[Y_e|A=a,W=w_i]$$

The simpler alternative is

$$\hat{\mathbb{E}}[Y_e|A=a] = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i 1[a_i=a]}{p(a)}$$

Also note $\mathbb{E}[Y_e|A=a,W] = \mathbb{E}[1[A=a]Y_e|W]/p(a)$.

Since

$$\mathbb{V}\mathrm{ar}[Y_e 1[A=a]] = \mathbb{E}[\mathbb{V}\mathrm{ar}[Y_e 1[A=a]|W]] + \mathbb{V}\mathrm{ar}\mathbb{E}[Y_e 1[A=a]|W]$$

we obtain $\mathbb{V}\mathrm{ar}[Y_e 1[A=a]] \geq \mathbb{V}\mathrm{ar}\mathbb{E}[Y_e 1[A=a]|W]$ which explains why conditioning on baseline W can be helpful.

Modeling of conditional expectation

If (Y_e, W) is jointly normal conditional on $A = a$, then
$\mathbb{E}[Y_e|A = a, W = w]$ is indeed a linear regression on a, w and wa.
If $\mathbb{E}[W|A = a] = 0$ then regression coefficient for A equals ψ.

In general the above is not true.

Nevertheless, as remarked by Emilie, we may still use least squares estimate from linear model!

In the following we elaborate a bit on this.

Best linear unbiased prediction

Consider a random variable Y and a random vector X. The best linear unbiased predictor (BLUP) of Y given X is

$$\hat{Y} = \mathbb{E}Y + \mathbb{C}\mathrm{ov}(Y, X)\mathbb{V}\mathrm{ar}(X)^{-1}(X - \mathbb{E}X) = \alpha^* + \phi^* X_c$$

with $X_c = X - \mathbb{E}X$, $\alpha^* = \mathbb{E}Y$ and
$\phi^* = \mathbb{C}\mathrm{ov}(Y, X)\mathbb{V}\mathrm{ar}(X)^{-1} = \mathbb{E}[YX_c]\mathbb{V}\mathrm{ar}(X)^{-1}$.

Let $Y = \hat{Y} + E$, where E is the prediction error. "Best" means that the variance of E is minimal among all linear predictors of Y and unbiased means $\mathbb{E}[Y - \hat{Y}] = \mathbb{E}E = 0$. In other words

$$(\alpha^*, \phi^*) = \underset{\alpha,\phi:\mathbb{E}[Y-\alpha-\phi X_c]=0}{\operatorname{argmin}} \mathbb{E}[Y - \alpha - \phi X_c]^2$$

Suppose we fit a (possibly misspecified) linear regression

$$Y_i = \alpha + \phi X_{ci} + \epsilon_i$$

for *iid* replications of (Y, X) with $X_{ci} = X_i - \bar{X}$.

Then by law of large numbers, least squares estimator

$$(\hat{\alpha}, \hat{\phi}) = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$$

(where $\mathbf{X}$ has rows $(1, X_{ci})$) converges to (α^*, ϕ^*)

That is, we consistently estimate the best linear approximation of the relation between Y and X !

Estimation of treatment effect

If we consider the case $Y = Y_e$ and $X = (A, W)$ then in the context of an RCT, the regression coefficient for A indeed estimates ψ.

See Appendix A in Højbjerre-Frandsen *et al.* (2025) or online book by Schuler.

Also it is straightforward to show that the estimator is asymptotically normal (see again Appendix A).

Remarkable: this holds even for misspecified linear model and non-normal data.

Choice of covariates

Regulatory guidelines specify that only a moderate number of covariates should be adjusted for. These should be specified prior to analysis.

Usually, baseline measurement of outcome variable are used.

Højbjerre-Frandsen *et al.* (2025) is concerned with construction of efficient adjustment variables based on historical data.

Further topics

Efficiency of various approaches to ATE estimates could be studied in further detail using the theory of efficient influence functions (e.g. paper by Schuler et al.)

Here I opted for a more direct and simple approach.

Summary

Considering solutions of exercises it can be seen that differencing, baseline-adjusting or mixed model do not provide any improvement if $\tau^2 = 0$ (in which case baseline and endpoint measurements are uncorrelated)

In absence of missing data, baseline adjusting and mixel model estimation can be expected to give similar results in terms of estimation variance

In case of missing data, mixed model estimation may be advantageous since it makes use of all observed data. In contrast, if a baseline measurement is missing, then the corresponding endpoint measurement can not be used in case of the baseline adjustment method (unless some kind of imputation is used which comes with further challenges)

We consider the issue of missing data in the next lecture.

Exercises

1. Recall the expression for a conditional distribution in a multivariate normal distribution and assume that the linear mixed model (1) is valid. Compute the joint covariance matrix of (Y_{aib}, Y_{aie}) and the conditional distribution of Y_{aie} given Y_{aib}. Compare with the linear model for baseline adjustment (2).
2. Assume that data are distributed according to the linear mixed model (1).
   - 2.1 Assume for simplicity that $\alpha_1 = \tau^2/(\sigma^2 + \tau^2)$ in (2) is known. What is then the estimate $\hat{\psi}_l$ of ψ ? (hint: you can simply base estimation on the differences $Y_{aie} - \alpha_1 Y_{aib}$)
   - 2.2 Compute the variances of $\hat{\psi}$, $\hat{\psi}_d$ and $\hat{\psi}_l$? Which estimate is optimal ?

Exercises continued

3. Assume for simplicity $\mu = \gamma = 0$ so that the design matrix X for the linear mixed model has just one column and that $n_c = n_T = n$. What is the variance of the weighted least squares estimate of ψ ?

   Hints: show first that the variance is of the form $(X^T \Sigma^{-1} X)^{-1}$. Further, each 2×2 block in Σ^{-1} is of the form $aE + bI$ for coefficients a and b where $E = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$.