Least squares and weighted least squares

Rasmus Waagepetersen

October 17, 2024

Least squares versus weighted least squares

Consider general linear mixed model (vector-matrix form):

$$Y = X\beta + ZU + \epsilon$$

Least squares estimate: $\hat{\beta} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y$

Linear mixed effects model estimate is *weighted least squares* estimate

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

where V is matrix of variances and covariances.

If data has nice structure (balanced), least squares (OLS) and weighted least squares (WLS) may coincide (e.g. orthodont data) but this is generally not the case.

Weighted least squares is superior to least squares (more accurate) because it takes into account correlations between observations.

Simple example: unbalanced one way ANOVA

$$Y_{ij} = \mu + U_i + \epsilon_{ij} \quad i = 1, 2$$

Variances $\tau^2 = 2$ and $\sigma^2 = 1$ so ICC is $2/3$.

For subject $i = 1$ only 1 observation. For subject $i = 2$, 4 observations - these are correlated and hence to some extent redundant.

Naive OLS estimate $\frac{1}{5}(Y_{11} + Y_{21} + Y_{22} + Y_{23} + Y_{24})$ - all observations same weight $1/5$.

Optimal WLS gives weight $3/7$ to Y_{11} and weights $1/7$ to remaining 4 observations.

Variance of naive estimate: 1.56
Variance of optimal estimate: 1.28.

Case study: fraction test results for 4th grade students

Computing with fractions is a key obstacle for primary school students

TRACK project is a randomized trial to test a new math teaching system against current practice.

125 schools were randomly allocated to current practice or new TRACK teaching system. 12717 students participate in the trial.

Currently we have data for Autumn 22 (baseline) and Spring 23 (one school year of treatment).

Fixed effects model

Parameters: main effect of teaching method (a) and main effect of participating in school year (b) Expected values in two-way table:

	A22 (baseline)	S23
Current	μ	$\mu + b$
New	μ	$\mu + b + a$

Which parameter is the key parameter ?

Least squares analysis (ordinary two-way ANOVA)

```
#create custom variable for treatment effect
> Treateffect=as.numeric(scores$Test==2 & scores$treatmentlabel=="TRACK
> 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 ?

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 (NB: does not make sense to add standard deviations)

Large correlation 0.72 between two test for same student

Now positive significant effect of intervention

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 random effects !

Difference for student with current:

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

Similarly, for student with new:

$$Y_{i2} - Y_{i1} = a + b + \tilde{\epsilon}_i$$

Least squares analysis based on differences

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

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

Residuals:
    Min      1Q  Median      3Q     Max
-52.664  -6.664  -0.336   6.664  60.336

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
```

Now treatment is effect of TRACK teaching method (similar estimate as in previous mixed model analysis.

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

Issue with difference approach: 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).

Mixed model analysis based on differences

We can still try adding random effect for class and school:

```
> fitdiffbrok=lmer(diffbrok~treatment.x+(1|instnr.x/Klasse.x),data=scoresdiff)
> summary(fitdiffbrok)
Random effects:
 Groups           Name        Variance Std.Dev.
 Klasse.x:instnr.x (Intercept) 12.454   3.529
 instnr.x          (Intercept)  3.262   1.806
 Residual                      99.305   9.965
Number of obs: 4498, groups:  Klasse.x:instnr.x, 249; instnr.x, 115

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)
(Intercept)  10.6798     0.4612 104.8952  23.155   <2e-16 *** #a
treatment.x   1.7133     0.6535 106.2772   2.622     0.01 *   #g
```

Class and school variances rather minor compared to residual variance

Standard errors higher than for linear regression due to estimated correlation within classes and schools. *p*-value for treatment bigger but still significant at 5% level.

Correlations at different level of hierarchy

Denote: σ^2, $\sigma^2_{student}$, σ^2_{class} and σ^2_{school} variances for noise, student, class and school.

Correlation between two tests for same student:

$$\frac{\sigma^2_{student} + \sigma^2_{class} + \sigma^2_{school}}{\sigma^2_{student} + \sigma^2_{class} + \sigma^2_{school} + \sigma^2}$$

(tests share all random effects except noise)

Correlation between two test for two students within the same class

$$\frac{\sigma^2_{class} + \sigma^2_{school}}{\sigma^2_{student} + \sigma^2_{class} + \sigma^2_{school} + \sigma^2}$$

(tests only share class and school random effects)

Correlation between two test for two students in different classes but same school

$$\frac{\sigma^2_{school}}{\sigma^2_{student} + \sigma^2_{class} + \sigma^2_{school} + \sigma^2}$$

(tests only share school random effect)

For last mixed model analysis based on differences there is no student effect. The two last correlations are

```
> (12.454+3.262)/(12.454+3.262+99.305)
[1] 0.1366359
> 3.262/(12.454+3.262+99.305)
[1] 0.02836004
```

Both correlations quite small - differencing may have worked.

Tests for variances ?

One may be interested in testing whether a variance is zero.

Unfortunately the theory for this is incomplete and it seems that no both generally and easily applicable approach is available.