Extending the model: 3 sources of variability
- Containers (j) randomly assigned to treatments (i)
  - Var containers = σ_c^2
- Plants (k) sampled within containers
  - Var plants(container) = σ_p^2
- Multiple (m) measurements on each plant
  - Var measurements(plants, containers) = σ_m^2

$$\begin{aligned} Y_{ijkm} &= \mu + \alpha_i + c_{ij} + p_{ijk} + \varepsilon_{ijkm} \\ c_{ij} &\sim N(0, \sigma_c^2) \\ p_{ijk} &\sim N(0, \sigma_p^2) \\ m_{ijkm} &\sim N(0, \sigma_m^2) \end{aligned}$$

What can go wrong?
- Variance components estimated to be **zero**
  - You expect variability between containers, but $\hat{\sigma_c^2} = 0$
- When $\hat{\sigma}_c^2 = 0$, that term dropped from the model
  - So barley analysis would become the wrong analysis
    - It assumes the eu = the plant (i.e., model without container effects)

Why does $\hat{\sigma}_c^2 \leq 0$ occur?
- Explanation clearer with Traditional approach - applies to REML
- Use barley study as a testbed: 3 plants per container, 2 containers per treatment

	$\hat{\sigma}_p^2$	Var container ave.		$\hat{\sigma}_c^2$
Real data	0.31	1.87	$= \frac{\sigma_c^2}{2} + \frac{0.31}{2\times3}$	3.64
Problem	2.00	0.20	$= \frac{\sigma_c^2}{2} + \frac{2.0}{2\times3}$	-0.27
	large	small		

- When container averages are "less variable than should be" $\Rightarrow$ est. VC < 0
  - REML forces it to be 0

Does a negative variance component even make sense?
- Variances are non-negative!
- But what if you think about correlations instead?

$$\text{ICC} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_p^2} = \frac{-0.27}{-0.27 + 2.00} = -0.16$$

  - That's a perfectly fine value for a correlation!
- But REML will force = 0 because a variance has to be non-negative

Does $\hat{\sigma}_c^2 = 0$ cause any problems?
- just said equivalent to ignoring containers

Here's a more detailed evaluation
Compare Traditional and REML estimates
Fact: the total variance, $\sigma_c^2 + \sigma_p^2$ remains approximately constant
Using the "Problem" data:

Method	total var.	$\hat{\sigma}_p^2$
Traditional	-0.27 + 2.00 = 1.73	2.00
REML	0 + ?? = 1.73	1.73

Consequences of dropping $\hat{\sigma}_p^2$ down
se for trt means and differences are wrong
could be too large or too small
wrong confidence intervals
tests have wrong type I error
You think you have $\alpha = 0.05$ tests or 95% ci's
but you don't

How do you know there is a problem?
Only signs are:
estimated variance component = 0
or error degrees of freedom are not what you expected

What does software let you do?
SAS
can specify nobound option that allows negative REML estimates
or use traditional method (no issues)
JMP
can specify traditional (EMS) estimation
R
enforces $\sigma^2 > 0$, no alternative
because internally it works with $\log \sigma^2$

Reasons why Traditional estimated VC < 0, with fixes

1. observations are negatively correlated
   possible reasons: competition, interference, dominance
   fixes: allow negative estimates (SAS, JMP)
   or recast the model in terms of correlation, (R, SAS, JMP)
2. you data has one or more outliers
   unusual value can increase $\hat{\sigma}_p^2$ a lot
   much less effect on $\hat{\sigma}_c^2 \Rightarrow$ "Problem" data
   fix: check for erroneous values, fix
3. you have the wrong model
   There are various ways to have a wrong model.
   One very common one is heterogeneous variances
   The model assumes σ_p^2 is the same in all containers
   When σ_p^2 is not constant, often $\Rightarrow$ neg. VC estimates

Check with a residual vs. predicted value plot
fixes: transform responses
or use a more complicated model with unequal variances

4. true $\sigma^2 = 0$
   there is a substantial probability that the estimate = 0
   bringing up to 0 and dropping the term are not problems

Vocabulary used to describe errors or issues

"not positive definite"

Concept of a PD variance-covariance matrix:
All variances are positive (not negative, not zero)
All correlations are between -1 and 1, none are exactly -1 or 1
Covariance = correlation * $\sqrt{\sigma_1^2 \times \sigma_2^2}$

Assemble the estimated variance components into a matrix
SAS calls this the G matrix, e.g.

$$G = \begin{bmatrix} \sigma_c^2 & 0 \\ 0 & \sigma_p^2 \end{bmatrix}$$

Most common cause: one of the estimated variance components = 0
or ≤ 0 if using the traditional method

See earlier discussion of estimated variance component = 0
Don't blindly ignore the warning.

"singular"

Two random effects in a model are perfectly correlated
Example: model has 3 levels of variability: container, plant, measurement
$Y_{ijkl} = \mu_i + c_{ij} + p_{ijk} + \varepsilon_{ijkl}$
i = treatment, j = container, k = plant, l = measurement
Fit to data with only one measurement per plant $\quad Y_{ijk} = \mu_i + c_{ij} + p_{ijk} + \varepsilon_{ijk}$
"plant" and "measurement" have the same subscripts!
data only informs you about their sum: $p_{ijk} + \varepsilon_{ijk}$, but not each part
can estimate $\sigma_p^2 + \sigma_e^2$ but not their sum

One curious thing that is not an error

non-integer degrees of freedom

A detailed explanation requires a long detour into expected mean squares
Short, semi-intuitive explanation

Some combinations of variance components can be estimated directly from the data
as sums-of-squared differences of various kinds
Can estimate $\sigma_c^2 + \sigma_p^2/3$ when all containers have 3 plants
These have integer degrees of freedom

When there are missing data, e.g., some containers have only 2 plants
Var $\overline{Y}_i$ may be equal to $\sigma_c^2/c + \sigma_p^2/(2.78\,c)$

Can calculate this from σ_c^2 and σ_p^2
But can not estimate as a sum of squared differences
So don't have a degrees of freedom

Best we can do is approximate the degrees of freedom
- Satterthwaite or Kenward-Roger approximations
- Both give non-integer degrees of freedom
- Many models, S and K-R give the same approximate df
- Kenward-Roger is more general, but slower to compute
- I use K-R as my default