STAT/BIOSTAT 571: Homework 7 Key

1. **[Estimating random effects variances]** The setting for this question is the 'unbalanced one-way random model', in which

$$\begin{aligned} a_i &\overset{i.i.d.}{\sim} N(0, \sigma_a^2), \text{ for } 1 \leq i \leq n \\ Y_{ij}|a_i &\overset{indept}{\sim} N(\beta_0 + a_i, \sigma_Y^2), \text{ for } 1 \leq j \leq n_i \end{aligned}$$

and where 'unbalanced' indicates that not all n_i are identical. By constructing two different unbiased estimates of σ_a^2 from this model's sufficient statistics, or otherwise, show that the sufficient statistics are not complete.

Note: this result shows that, even without a complex form of LMM, there is no uniformly minimum variance unbiased estimator of σ_a^2. Consequently, in analysis of finite samples there will always be *some* room for debate about what constitutes the "best" estimate of this parameter – or any related parameter.

Solution. We proceed by first deriving the sufficient statistics and then finding a non-zero function of the statistics that is unbiased for zero.

To find the sufficient statistics, we must express the joint multivariate distirbution, which involves the inverse and determinant of $\boldsymbol{\Sigma}$, which is of the form

$$\boldsymbol{\Sigma} = \sigma_Y^2 \boldsymbol{I}_{n_i} + \sigma_a^2 \boldsymbol{J}_{n_i},$$

where $\boldsymbol{J}_{n_i}$ is the $n_i \times n_i$ matrix of ones. Applying the Sherman-Morrison-Woodbury formula, we write

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_Y^2}\left(\boldsymbol{I}_{n_i} - \frac{\sigma_a^2}{\sigma_Y^2 + n_i\sigma_a^2}\boldsymbol{J}_{n_i}\right).$$

Now, the joint density function of the $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{Y}|\beta_0, \sigma_Y^2, \sigma_a^2) = \prod_{i=1}^{n}(2\pi)^{-n_i/2}|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i})^T\boldsymbol{\Sigma}_i^1(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i})\right\}.$$

Since this follows an exponential family, we can focus our attention on the exponent, rewriting it as

$$\begin{aligned} &\sum_{i=1}^{n}(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i})^T\boldsymbol{\Sigma}_i^1(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i}) \\ =&\sum_{i=1}^{n}(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i})^T\frac{1}{\sigma_Y^2}\left(\boldsymbol{I}_{n_i} - \frac{\sigma_a^2}{\sigma_Y^2 + n_i\sigma_a^2}\boldsymbol{J}_{n_i}\right)(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i}) \\ =&\frac{1}{\sigma_y^2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i})^T(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i}) - \frac{\sigma_a^2}{\sigma_Y^2}\sum_{i=1}^{n}\frac{1}{\sigma_Y^2 + n_i\sigma_a^2}(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i})^T\boldsymbol{J}_{n_i}(\boldsymbol{Y}_i - \beta_0\mathbf{1}_{n_i}) \\ =&\frac{1}{\sigma_Y^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n_i}Y_{ij}^2\right] - \frac{2\beta_0}{\sigma_Y^2}\sum_{i=1}^{n}\left[\sum_{j=1}^{n_i}Y_{ij}\right] + \frac{\beta_0}{\sigma_Y^2}\sum_{i=1}^{n}n_i \end{aligned}$$

$$+\frac{\sigma_a^2}{\sigma_Y^2}\sum_{i=1}^{n}\frac{1}{\sigma_Y^2+n_i\sigma_a^2}\left[\sum_{j=1}^{n_i}\sum_{k=1}^{n_i}Y_{ij}Y_{ik}\right]-\frac{2\beta_0\sigma_a^2}{\sigma_Y^2}\sum_{i=1}^{n}\frac{1}{\sigma_Y^2+n_i\sigma_a^2}\left[\sum_{j=1}^{n_i}Y_{ij}\right]+\frac{\beta_0^2\sigma_a^2}{\sigma_Y^2}\sum_{i=1}^{n}\frac{n_i^2}{\sigma_Y^2+n_i\sigma_a^2}.$$

From here, we see that sufficient statistics are

$$\begin{aligned}
T_1(\boldsymbol{Y}) &= \sum_i\sum_j Y_{ij}^2, \quad 1\le i\le n, \quad 1\le j\le n_i;\\
T_{2,i}(\boldsymbol{Y}) &= \sum_j\sum_k Y_{ij}Y_{ik}, \quad 1\le i\le n, \quad 1\le j,k\le n_i;\\
T_{3,i}(\boldsymbol{Y}) &= \sum_j Y_{ij}, \quad 1\le i\le n, \quad 1\le j\le n_i.
\end{aligned}$$

We note that the following expecations hold:

$$\begin{aligned}
E[T_1(\boldsymbol{Y})] &= (\sigma_Y^2+\sigma_a^2+\beta_0^2)\sum_{i=1}^{n}n_i\\
&= (\sigma_Y^2+\sigma_a^2+\beta_0^2)N;\\
E[T_{2,i}(\boldsymbol{Y})] &= n_i^2(\sigma_a^2+\beta_0^2), \quad 1\le i\le n;\\
E[T_{3,i}^2(\boldsymbol{Y})] &= n_i\sigma_Y^2+n_i^2(\sigma_a^2+\beta_0^2), \quad 1\le i\le n.
\end{aligned}$$

This means that we can construct unbiased estimates of σ_Y^2 (or other parameters) in multiple ways. For instance,

$$\begin{aligned}
\hat{\sigma}_Y^2 &= \frac{1}{N}T_1-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i^2}T_{2,i};\\
\tilde{\sigma}_Y^2 &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}(T_{3,i}^2-T_{2,i}).
\end{aligned}$$

Hence, $E[\hat{\sigma}_Y^2-\tilde{\sigma}_Y^2]=0$ and $Pr(\hat{\sigma}_Y^2-\tilde{\sigma}_Y^2=0)<1$, so the sufficient statistics for this model are not complete.

2. **[LMMs and REML]** This question closely follows Q8.2 from Jon's book. The `dyestuff` data come from a balanced study of the yield from six randomly chosen batches of raw material, with five replicates each. The aim of this experiment was to find out whether batch-to-batch variation was responsible for significant variation in the final product yield.

   We will assume the one-way random effects model;

$$\begin{aligned}
a_i &\overset{i.i.d}{\sim} N(0,\sigma_a^2)\\
Y_{ij}|a_i &\overset{indept}{\sim} N(\mu+a_i,\sigma_\epsilon^2)
\end{aligned}$$

   for $1\le i\le n=6$ batches, each with $1\le j\le n_i=5$ replicates. (An alternative statement of the second line is that $Y_{ij}=\mu+a_i+\epsilon_{ij}$ with all $\epsilon_{ij}\overset{i.i.d}{\sim}N(0,\sigma_\epsilon^2)$.)

   (a) State the marginal distribution of the $\{Y_{ij}\}$, integrating out the a_i.

Solution. By convolution properties of the normal distribution, we know that

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T \sim MVN_{n_i}(\mu_i, \boldsymbol{\Sigma}_i).$$

Hence we need only find the first two moments of the distribution. We calculate μ_i and the elements of $\boldsymbol{\Sigma}_i$ as follows:

$$\begin{aligned}
\mathrm{E}[Y_{ij}] &= \mathrm{E}[\mathrm{E}[Y_{ij}|a_i]] \\
&= \mathrm{E}[\mu + a_i] \\
&= \mu \\
\mathrm{Var}[Y_{ij}] &= \mathrm{Var}[\mathrm{E}[Y_{ij}|a_i]] + \mathrm{E}[\mathrm{Var}[Y_{ij}|a_i]] \\
&= \mathrm{Var}[\mu + a_i] + \mathrm{E}[\sigma_\epsilon^2] \\
&= \sigma_a^2 + \sigma_\epsilon^2 \\
\mathrm{Cov}[Y_{ij}, Y_{ik}] &= \mathrm{Cov}[\mathrm{E}[Y_{ij}|a_i], \mathrm{E}[Y_{ik}|a_i]] + \mathrm{E}[\mathrm{Cov}[Y_{ij}, Y_{ik}|a_i]] \\
&= \mathrm{Cov}[\mu + a_i, \mu + a_i] + E[0] \\
&= \mathrm{Cov}[a_i, a_i] \\
&= \mathrm{Var}[a_i] \\
&= \sigma_a^2
\end{aligned}$$

(b) State a relevant null hypothesis in two ways; i) in terms of the a_i ii) in terms of σ_a^2.

Solution. A relevant null hypothesis is that there are no batch effects in the experiment. For our assumed model, this can be expressed in terms of the values of the random effects or their variance, defined respectively as H_{0,a_i} and H_{0,σ_a^2}, with

$$H_{0,a_i} : a_1 = a_2 = a_3 = a_4 = a_5 = a_6 = 0,$$

$$H_{0,\sigma_a^2} : \sigma_a^2 = 0.$$

(c) Show that the MLE for μ is $\bar{Y}_{\cdot\cdot} = n^{-1} \sum_i n_i^{-1} \sum_j Y_{ij}$, the 'grand mean' of the yields.

Solution. From (a), we can write the log-likelihood

$$\log L(y|\mu, \sigma_a^2 \sigma_\epsilon^2) = \log \prod_{i=1}^{n} \left[(2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{ \frac{-1}{2} (\boldsymbol{y}_i - \mu \mathbf{1}_{n_i})^T \Sigma^{-1} (\boldsymbol{y}_i - \mu \mathbf{1}_{n_i}) \right\} \right].$$

Recall that Σ is an $n_i \times n_i$ matrix with $\sigma_a^2 + \sigma_\epsilon^2$ on the diagonals and σ_a^2 on the off-diagonals. We can simplify the above expression to

$$\begin{aligned}
\log L(y|\mu, \sigma_a^2 \sigma_\epsilon^2) = &-\frac{n n_i}{2} \log 2\pi - \frac{n_i - 1}{2} \log \sigma_\epsilon^2 - \frac{n_i}{2} \log(\sigma_\epsilon^2 + n_i \sigma_a^2) \\
&- \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2 + \frac{n_i^2 \sigma_a^2}{2\sigma_\epsilon^2 (\sigma_\epsilon^2 + n_i \sigma_a^2)} \sum_{i=1}^{n} (\bar{y}_{i\cdot} - \mu)^2.
\end{aligned}$$

We then have the score equation

$$\frac{\partial \log L}{\partial \mu} = \frac{n_i n}{\sigma_\epsilon^2} (\bar{y}_{\cdot\cdot} - \mu) - \frac{n_i n \sigma_a^2}{\sigma_\epsilon^2 (\sigma_\epsilon^2 + n_i \sigma_a^2)} (\bar{y}_{\cdot\cdot} - \mu) = 0,$$

which is solved by the grand mean $\hat{\mu}_{MLE} = \bar{y}_{\cdot\cdot}$ as desired.

(d) As seen in class, the REML estimates for variance terms in linear mixed models minimize

$$(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + \log|\boldsymbol{V}| + \log|\mathbf{X}^T \boldsymbol{V}^{-1} \mathbf{X}|$$

where $\boldsymbol{Y}$ denotes the $nn_i \times 1$ vector of all outcomes, $\boldsymbol{V}$ denotes the (marginal) variance of $\boldsymbol{Y}$, $\mathbf{X}$ is the $nn_i \times p$ design matrix, and $\boldsymbol{\beta}$ is the usual length p vector of fixed effects. For the one-way model, give formula for the REML estimates for all three parameters; μ, σ_a^2 and σ_ϵ^2 (Hint: revise/look up the Sherman-Morrison-Woodbury formula and the matrix determinant lemma.)

Solution. From the expression for REML given above, β is μ from our model, $\mathbf{X}$ is $\mathbf{1}_{nn_i}$ (a vector of nn_i ones), $\boldsymbol{Y}$ is the $nn_i \times 1$ vector of stacked outcomes, and $\mathbf{V}$ is the $nn_i \times nn_i$ block diagonal matrix given by

$$\mathbf{V} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma} \end{bmatrix},$$

with $\boldsymbol{\Sigma}$ as defined in parts (a) and (c), and $\mathbf{0}$ the $n_i \times n_i$ matrix of zeros. Applying the Sherman-Morrison-Woodbury formula, we write

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_\epsilon^2} \left(\boldsymbol{I}_{n_i} - \frac{\sigma_a^2}{\sigma_\epsilon^2 + n_i \sigma_a^2} \boldsymbol{J}_{n_i} \right)$$

where $\boldsymbol{J}_{n_i}$ is the $n_i \times n_i$ matrix of ones. It follows that

$$\mathbf{V}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}^{-1} \end{bmatrix}.$$

For the determinant $|\mathbf{V}|$, we note that $|\mathbf{V}| = |\boldsymbol{\Sigma}|^n$. Applying the matrix determinant lemma for $\boldsymbol{\Sigma}$, we find

$$\begin{aligned} \boldsymbol{\Sigma} &= \sigma_\epsilon^2 \boldsymbol{I}_{n_i} + (\sigma_a \mathbf{1}_{n_i})(\sigma_a \mathbf{1}_{n_i})^T \\ \Rightarrow |\boldsymbol{\Sigma}| &= \left(1 + \frac{1}{\sigma_\epsilon^2} (\sigma_a \mathbf{1}_{n_i})^T (\sigma_a \mathbf{1}_{n_i}) \right) (\sigma_\epsilon^2)^{n_i} \\ &= (\sigma_\epsilon^2)^{n_i - 1} (\sigma_\epsilon^2 + n_i \sigma_a^2) \\ \Rightarrow |\mathbf{V}| &= (\sigma_\epsilon^2)^{(n_i - 1)n} (\sigma_\epsilon^2 + n_i \sigma_a^2)^n. \end{aligned}$$

Denoting the REML objective function by ℓ, we use our matrix results to simplify and minimize ℓ:

$$\begin{aligned} \ell &= (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}) + \log|\mathbf{V}| + \log|\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}| \\ &= \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}) + \log(\sigma_\epsilon^2)^{(n_i - 1)n} (\sigma_\epsilon^2 + n_i \sigma_a^2)^n + \log \left| \sum_{i=1}^{n} \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1} \right| \\ &= \sum_{i=1}^{n} \left[\sum_{j=1}^{n_i} \frac{(y_{ij} - \bar{y}_{i\cdot})^2}{\sigma_\epsilon^2} + \frac{n_i}{\sigma_\epsilon^2 + n_i \sigma_a^2} (\bar{y}_{i\cdot} - \mu)^2 \right] + n(n_i - 1) \log \sigma_\epsilon^2 + (n - 1) \log(\sigma_\epsilon^2 + n_i \sigma_a^2) + \log(nn_i). \end{aligned}$$

$$\frac{\partial \ell}{\partial \sigma_\epsilon^2} = \frac{-1}{(\sigma_\epsilon^2)^2} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left[(y_{ij} - \bar{y}_{i\cdot})^2 - \frac{1}{(\sigma_\epsilon^2 + n_i \sigma_a^2)^2} n_i (\bar{y}_{i\cdot} - \mu)^2 \right] + \frac{n(n_i - 1)}{\sigma_\epsilon^2} + \frac{n-1}{\sigma_\epsilon^2 + n_i \sigma_a^2},$$

$$\frac{\partial \ell}{\partial \sigma_a^2} = -\sum_{i=1}^{n} (\bar{y}_{i\cdot} - \mu)^2 \frac{n_i^2}{(\sigma_\epsilon^2 + n_i \sigma_a^2)^2} + \frac{(n-1) n_i}{\sigma_\epsilon^2 + n_i \sigma_a^2}.$$

Setting the partial derivatives equal to zero, we find that the estimating equations are solved by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{n(n_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

$$\hat{\sigma}_a^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_{i\cdot} - \hat{\mu})^2 - \frac{1}{n_i} \hat{\sigma}_\epsilon^2.$$

(e) In the one-way random effects model with balanced data, it can be shown that

$$\frac{\frac{n_i}{n-1} \sum_i (Y_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 / (n_i \sigma_a^2 + \sigma_\epsilon^2)}{\frac{1}{n(n_i - 1)} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2 / \sigma_\epsilon^2} \sim F_{n-1, n(n_i-1)}.$$

Using this result, describe how to obtain a p-value assessing the null hypothesis in b). Implement your method for the dyestuff data.

Solution. Under the null hypothesis, $\sigma_a^2 = 0$. This, with the result above, yields the statistic

$$T_n := \frac{\frac{n_i}{n-1} \sum_i (Y_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\frac{1}{n(n_i-1)} \sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2} \sim F_{n-1, n(n_i-1)}.$$

We thus obtain a p-value by evaluating $p = \Pr\left(F_{n-1,n(n_i-1)} > T_n\right)$. In the `dyestuff` data, we find $p = 0.0044$. The `R` code for this can be found in the appendix.

Hint: to check your formulae, code them 'by hand' in `R` and compare them against `lme()` output on the dyestuff data.

(f) Using the same result as in e), show that the probability the REML estimate of σ_a^2 is negative is given by the upper tail of $F_{n(n_i-1),(n-1)}$, beyond $1 + n_i \sigma_a^2 / \sigma_\epsilon^2$. What does `lme()` do when it encounters these negative variance estimates?

Solution. We first express the event $\{\hat{\sigma}_a^2 < 0\}$ as follows:

$$\begin{aligned}
\{\hat{\sigma}_a^2 < 0\} &= \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_{i\cdot} - \hat{\mu})^2 - \frac{1}{n_i n(n_i - 1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2 < 0 \right\} \\
&= \left\{ \frac{n n_i (n_i - 1)}{n-1} < \frac{\sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{\sum_{i=1}^{n} (\bar{y}_{i\cdot} - \hat{\mu})^2} \right\} \\
&= \left\{ n_i < \frac{\frac{1}{n(n_i-1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2}{\frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_{i\cdot} - \hat{\mu})^2} \right\}
\end{aligned}$$

$$= \left\{ n_i \frac{\sigma_a^2 + \sigma_\epsilon^2}{\sigma_\epsilon^2} < \frac{\frac{1}{n(n_i-1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 / \sigma_\epsilon^2}{\frac{n_i}{n-1} \sum_{i=1}^{n} (\bar{y}_{i.} - \hat{\mu})^2 / (n_i \sigma_a^2 + \sigma_\epsilon^2)} \right\}$$
$$= \left\{ 1 + n_i \frac{\sigma_a^2}{\sigma_\epsilon^2} < \frac{\frac{1}{n(n_i-1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 / \sigma_\epsilon^2}{\frac{n_i}{n-1} \sum_{i=1}^{n} (\bar{y}_{i.} - \hat{\mu})^2 / (n_i \sigma_a^2 + \sigma_\epsilon^2)} \right\}$$

Hence,

$$\Pr\left(\hat{\sigma}_a^2 < 0\right) = \Pr\left(1 + n_i \frac{\sigma_a^2}{\sigma_\epsilon^2} < \frac{\frac{1}{n(n_i-1)} \sum_{i=1}^{n} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 / \sigma_\epsilon^2}{\frac{n_i}{n-1} \sum_{i=1}^{n} (\bar{y}_{i.} - \hat{\mu})^2 / (n_i \sigma_a^2 + \sigma_\epsilon^2)} \right).$$

We know from probability theory that $X \sim F_{v,u}$ implies $\frac{1}{X} \sim F_{u,v}$, which completes the proof.

3. **[Using random slopes without clustering]**

   (a) Explain how for *independent* outcomes, LMM software fitting random slopes can be used to fit heteroskedastic linear models, i.e. how to find MLEs for all the parameters in

   $$Y_i | \boldsymbol{X}_i, Z_i \sim N(\boldsymbol{X}_i^T \boldsymbol{\beta}, \sigma_{Z_i}^2),$$

   where $\boldsymbol{X}_i$ is a vector of covariates and Z_i is a single, categorical covariate

   Solution. For independent outcomes, LMM software can fit heteroskedastic linear models, obtaining the MLEs for the parameters in

   $$Y_i | \boldsymbol{X}_i, Z_i \sim N(\boldsymbol{X}_i^T \boldsymbol{\beta}, \sigma_{Z_i}^2).$$

   This is obtained by postulating the mixed model

   $$\boldsymbol{Y} | \mathbf{X}, \mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}b + \boldsymbol{\epsilon},$$
   $$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_n),$$
   $$b \sim N_n(\mathbf{0}, \mathrm{diag}[\sigma_{Z_1}^2 - \sigma_\epsilon^2, \ldots, \sigma_{Z_n}^2 - \sigma_\epsilon^2]).$$

   Now, we have that

   $$Y_i | \boldsymbol{X}_i, Z_i, b_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + Z_i b_i + \epsilon_i,$$

   and hence the marginal distribution (with respect to b only) corresponds to the heteroskedastic linear model

   $$Y_i | \boldsymbol{X}_i, Z_i \sim N(\boldsymbol{X}_i^T \boldsymbol{\beta}, \sigma_{Z_i}^2).$$

   We obtain the MLE of the coefficients in the heteroskedastic model by using LMM software in an "off the shelf" manner provided we include random slopes for the categorical variable Z_i (note that there is no random intercept in this LMM!). We are also interested in obtaining the MLEs for each of the variances, $\sigma_{Z_i}^2$, from the LMM software. Recall that for the k-th category of Z_i, the MLE of the variance is

   $$\hat{\sigma}_k^2 = \frac{1}{n_k} \sum_{j=1}^{n_k} (Y_{kj} - \boldsymbol{X}_{kj}^T \hat{\boldsymbol{\beta}})^2,$$

   where $n_k = \sum_{i=1}^{n} \mathbf{1}[Z_i = k]$, with Y_{kj} and $\boldsymbol{X}_{kj}$ are the j-th outcome and covariate vector in category k. Hence, we can calculate the MLEs of the $\sigma_{Z_i}^2$ from the LMM's fixed effect residuals.

   (b) Using an example (of your own devising) based on R's built-in `cars` dataset, illustrate how the two-sample t-test with unequal variances (i.e. the default approach using `t.test()`) is essentially just a Wald test from the fit in a). Note: As well as explaining your example, your answer should explain the exact relationship between these two procedures

Solution. There are many possible ways to split the data into two groups for testing. One option is to split the data by `speed > 15` and `speed <= 15`, testing the difference in the mean `dist` between the two groups. This is equivalent to the model

$$Y_i|Z_i, b_i = \mu_0 + Z_i(\mu_1 - \mu_0 + b_i) + \epsilon_i = \beta_0 + Z_i(\beta_1 + b_i) + \epsilon_i = \boldsymbol{Z}^T\boldsymbol{\beta} + Z_i b_i + \epsilon_i,$$

where Z_i is an indicator of `speed > 15`. For this data, we conduct the t-test for unequal variances and Wald test from LMM fit by maximum-likelihood, as well as the Wald test from LMM fit by REML.

Method	`t.test()`	`lmer(,REML=FALSE)`	`lmer(,REML=TRUE)`
Statistic	-5.582737	5.699493	5.582737

If we compare the default behavior of `t.test()` to `lmer()` with `REML=FALSE`, we note that the Wald statistic differs slightly from the t-statistic (the opposite signs result from taking the square root of a quadratic form). If we set `REML=TRUE` for fitting the LMM, we obtain equality of the Wald statistic and squared t-statistic. This occurs because `t.test()` is based on unbiased estimates of the sample variance, which are equal to $n_k/(n_k-1)\hat{\sigma}_k^2$, a multiple of the MLEs $(\hat{\sigma}_0^2, \hat{\sigma}_1^2)$ we obtain from fitting the LMM.

To demonstrate the reason for equality of the Wald test and t-test when we optimize the REML criterion, note that our variance parameter estimates are chosen to minimize

$$\begin{aligned}
REML(\boldsymbol{\beta}, \sigma_0^2, \sigma_1^2) &= \log|\mathbf{V}| + \log|\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{Z}| + (\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{Z}\boldsymbol{\beta}) \\
&= \sum_{i=1}^n \log\sigma_{Z_i}^2 + \sum_{i=1}^n \sigma_{Z_i}^{-2}(Y_i - \beta_0 - z_i\beta_1)^2 \\
&+ \log\left\{\left(\sum_{i=1}^n z_i^2\sigma_{Z_i}^{-2}\right)\left(\sum_{i=1}^n \sigma_{Z_i}^{-2}\right) - \left(\sum_{i=1}^n z_i\sigma_{Z_i}^{-2}\right)^2\right\} \\
&= \sum_{i=1}^n \log\sigma_{Z_i}^2 + \sum_{i=1}^n \sigma_{Z_i}^{-2}(Y_i - \beta_0 - z_i\beta_1)^2 + \log\left\{\sum_{i=1}^n\sum_{j=1}^n \sigma_{Z_i}^{-2}\sigma_{Z_j}^{-2} z_j(z_j - z_i)\right\} \\
&= \sum_{i=1}^n \log\sigma_{Z_i}^2 + \sum_{i=1}^n \sigma_{Z_i}^{-2}(Y_i - \beta_0 - z_i\beta_1)^2 + \log\left\{n_{01}\sigma_0^{-2}\sigma_1^{-2}\right\} \\
&= (n_0\log\sigma_0^2 + n_1\log\sigma_1^2) + \sum_{i=1}^n \sigma_{Z_i}^{-2}(Y_i - \beta_0 - z_i\beta_1)^2 + (\log n_{01} - \log\sigma_0^2 - \log\sigma_1^2) \\
&= (n_0 - 1)\log\sigma_0^2 + (n_1 - 1)\log\sigma_1^2 + \sum_{i=1}^n \sigma_{Z_i}^{-2}(Y_i - \beta_0 - z_i\beta_1)^2 + \log n_{01} \\
&= (n_0 - 1)\log\sigma_0^2 + \sum_{j=1}^{n_0} \sigma_0^{-2}(Y_{0i} - \beta_0)^2 + \log n_{01} \\
&+ (n_1 - 1)\log\sigma_1^2 + \sum_{j=1}^{n_1} \sigma_0^{-2}(Y_{1i} - \beta_0 - \beta_1)^2
\end{aligned}$$

where $\mathbf{V} = \text{diag}(\sigma_{Z_1}^2, \ldots, \sigma_{Z_n}^2)$ and n_{01} is the number of pairs $(Z_i, Z_j) = (0, 1)$. Recalling that for this model, $\hat{\boldsymbol{\beta}} = (\bar{Y}_0, \bar{Y}_1 - \bar{Y}_0)^T$, we can see that the REML criterion is minimized by the

unbiased sample variance estimates, s_0^2 and s_1^2, with

$$s_k^2 = \frac{1}{n_k - 1}\sum_{j=1}^{n_k}(Y_{ki} - \hat{\beta}_0 - k\hat{\beta}_1)^2 = \frac{1}{n_k - 1}\sum_{j=1}^{n_k}(Y_{ki} - \bar{Y}_0 - k(\bar{Y}_1 - \bar{Y}_0))^2, \quad k \in \{0, 1\}.$$

Now, the Wald test of $H_0 : \beta_1 = 0$ uses the statistic

$$W = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \quad (\text{or } W^2),$$

where we estimate the standard error based on the variance

$$\begin{aligned}
\text{Var}(\hat{\beta}|\mathbf{Z}) &= \left(\sum_{i=1}^{n} \boldsymbol{Z}_i \sigma_{Z_i}^{-2} \boldsymbol{Z}_i^T\right)^{-1} \\
&= \left(\frac{1}{\sigma_0^2}\begin{pmatrix} n_0 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\sigma_1^2}\begin{pmatrix} n_1 & n_1 \\ n_1 & n_1 \end{pmatrix}\right)^{-1} \\
&= \begin{pmatrix} \frac{\sigma_0^2}{n_0} & -\frac{\sigma_0^2}{n_0} \\ -\frac{\sigma_0^2}{n_0} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}
\end{aligned}$$

We note that a consistent estimator for $\text{Var}(\hat{\beta}|\mathbf{Z})$ is obtained by plugging in either the MLE or REML estimates for σ_0^2 and σ_1^2, leaving us with either

$$\hat{SE}(\hat{\beta}_1)_{REML} = \sqrt{\frac{\hat{\sigma}_0^2}{n_0} + \frac{\hat{\sigma}_1^2}{n_1}} \quad \text{or} \quad \hat{SE}(\hat{\beta}_1)_{MLE} = \sqrt{\frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}},$$

which are the estimators used in `lmer()`. This explains why `t.test()` and the Wald test based on the fit returned from `lmer(, REML=FALSE)` differ, with the t-statistic slightly smaller than the Wald statistic, while `lmer(, REML=TRUE)` (the default argument!) yields an equivalent statistic to `t.test()`. Note that since the variance estimates are consistent in both cases, this discrepancy becomes negligible as sample size increases.

Appendix: R code

```
#### Question 2e
dye <- read.table("dyestuff.txt", header=T)
y.. <- mean(dye$yield) # grand mean
yi. <- by(dye$yield, dye$batch, mean) # cluster means
num <- (5/5)*sum((yi.-y..)^2) # numerator of test statistic
den <- (1/(6*4))*sum((dye$yield-rep(yi., each=5))^2) # denom of test stat
1-pf(num/den, df1=5, df2=24) # p-value

#### Question 3b
library(lme4)
data(cars)
cars$fast <- (cars$speed > 15) * 1
cars$id <- 1:50
cars$fast.id <- with(cars, 1 * fast * id)

t.test.fit <- t.test(dist ~ fast, data = cars)
lmer.ml <- lmer(dist ~ fast + (fast - 1 | fast.id), data = cars, REML = FALSE)
lmer.reml <- lmer(dist ~ fast + (fast - 1 | fast.id), data = cars, REML = TRUE)

res <- c(
  Welch = t.test.fit$statistic,
  Wald.REML = summary(lmer.reml)$co["fast", "t value"],
  Wald.ML = summary(lmer.ml)$co["fast", "t value"]
)

res
```