Exam style questions from the review lecture.

Note that the length of this set of questions does not necessarily reflect the length of the exam.

1. Based around the nested random effects model

$$y_{ijk} = \alpha_i + \gamma_{ij} + e_{ijk},\ i = 1, \ldots, a,\ j = 1, \ldots, g,\ k = 1, \ldots, r$$

with $\gamma_{ij} \sim N(0, \sigma_g^2)$, $e_{ijk} \sim N(0, \sigma_e^2)$. [Example: α_i is breed of sheep, γ_{ij} is the jth sheep in breed i, and we measure the wool obtained from each sheep k times.]

   (a) What is $\mathrm{cov}(y_{ijk}, y_{i'j'k'})$?

$$\mathrm{cov}(y_{ijk}, y_{i'j'k'}) = \begin{cases} \sigma_g^2 + \sigma_e^2 & \text{if } i = i',\ j = j',\ k = k' \\ \sigma_g^2 & \text{if } i = i',\ j = j',\ k \neq k' \\ 0 & \text{otherwise} \end{cases}$$

   (b) We can write $SSA = \sum_i (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$. Find and expression for its expectation in terms of the α_i, σ_e^2 and σ_g^2.
   We have that

$$\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot} = (\alpha_i - \bar{\alpha}) + (\bar{\gamma}_{i\cdot} - \bar{\gamma}_{\cdot\cdot}) + (\bar{\epsilon}_{i\cdot\cdot} - \bar{\epsilon}_{\cdot\cdot\cdot})$$

   since the terms inside the brackets are independent (so their cross terms have expectation zero), we can write the expected sum of squares as

$$ESSA = \sum (\alpha_i - \bar{\alpha})^2 + E\bar{\gamma}^T C_a \bar{\gamma} + E\bar{\epsilon}^T C_a \bar{\epsilon}$$

   where

$$\bar{\gamma} = (\bar{\gamma}_{1\cdot}, \ldots, \bar{\gamma}_{a\cdot}) \sim N(0, \sigma_g^2 / g I_a)$$
$$\bar{\epsilon} = (\bar{\epsilon}_{1\cdot\cdot}, \ldots, \bar{\epsilon}_{a\cdot\cdot}) \sim N(0, \sigma_e^2 / gr I_a)$$

   hence

$$ESSA = \sum \alpha_i^2 + (a-1)\sigma_g^2/g + (a-1)\sigma_e^2/gr.$$

Alternatively, we could argue that

$$\bar{y}_{i\cdot\cdot} \sim N(\alpha_i, \sigma_g^2/g + \sigma^2 e/gr)$$

independently, hence

$$SSA = \bar{\mathbf{y}}^T C_a \bar{\mathbf{y}}$$

with the same $ESSA$ as above.

(c) Show that $MSG = \frac{1}{a(g-1)} \sum_{ij} (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$ has the same expectation as $MSA = SSA/g$ under the null hypothesis that $\alpha_1 = \ldots = \alpha_a = \bar{\alpha}$.
Here we observe that

$$\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} = \gamma_{ij} - \bar{\gamma}_{i\cdot} + \bar{\epsilon}_{ij\cdot} - \bar{\epsilon}_{i\cdot\cdot} = C_g \mathbf{z}_i$$

if

$$\mathbf{z}_i = (\gamma_{i1} + \bar{\epsilon}_{i1\cdot}, \ldots, \gamma_{ig} + \bar{\epsilon}_{ig\cdot})^T \sim N(0, \sigma_g^2 + \sigma_e^2/r)$$

so that

$$ESSG = E \sum_i \mathbf{z}_i^T C_g \mathbf{z}_i = a(g-1)(\sigma_g^2 + \sigma_e^2/r).$$

Hence,

$$EMSG = \sigma_g^2 + \sigma_e^2/r = gESSA/(a-1)$$

when the $\alpha_i = \bar{\alpha}$.

(d) We would like to formally show that we get an F test out of $gMSA/MSG$ to do this, we'll write $z_{ij} = \bar{y}_{ij\cdot}$.

i. Give an expression for z in vector form; show that its covariance can be written as $\tau^2 I$ (and find τ).
We note that the $\bar{y}_{ij\cdot} = \alpha_i + \gamma_{ij} + \bar{e}_{ij\cdot}$ and the random terms $\gamma_{ij} + \bar{e}_{ij\cdot}$ are independent over different ij hence, all the z_{ij} are independent with variance

$$\tau^2 = \sigma_g^2 + \sigma_e^2/r$$

ii. Write SSA and SSG as $z^T A_1 z$ and $z^T A_2 z$ with A_1 and A_2 idempotent and $A_1 A_2 = 0$.
Here we see that $\bar{y}_{i\cdot\cdot}$ are the fitted values from regressing $\mathbf{z}$ on the indicator functions for the α_i. So that

$$SSA = \mathbf{z}^T HCH\mathbf{z}$$

moreover,

$$SSG = \mathbf{z}^T (I - H)\mathbf{z}$$

and we know that $HCH(I - H) = 0$.
Alternatively, $A_1 = C_a \otimes \bar{J}_g$ and $A_2 = (I_a \otimes C_g)$ and we see that since all terms in the Kronecker products are idempotent, so are A_1 and A_2.
Further $A_1 A_2 = (C_a \otimes \bar{J}_g C_g) = 0$.

iii. Hence, show that $gMSA/MSG$ has an F distribution. Give its degrees of freedom.
From ii, $SSA/(\sigma_g^2 + \sigma_e^2/r) \sim \chi^2_{(a-1)}$ *and* $SSG(\sigma_g^2 + \sigma_e^2/r) \sim \chi^2_{g(a-1)}$ *and these are independent, so* $gMSA/MSG \sim F^{(a-1)}_{g(a-1)}$.

iv. Bonus: when H_0 is not true, what is the noncentrality parameter in the distribution above?
In SSA we have that the noncentrality term is $\sum(\alpha_i - \bar{\alpha})^2 = \theta_a$ *hence*

$$\frac{1}{\sigma_g^2 + \sigma_e^2/r} SSA \sim \chi^2_{a-1}(g\theta_a/(\sigma_g^2 + \sigma_e^2/r))$$

and thus the noncentrality parameter is

$$\lambda = g \sum (\alpha_i - \bar{\alpha})^2/(\sigma_g^2 + \sigma_e^2/r)$$

2. Contrasts. Here we assume that α has four levels and we are using reference coding.

   (a) Find a contrast matrix to test the hypotheses: $\alpha_1 = \alpha_2,\ \alpha_3 = (\alpha_1 + \alpha_2)/2,\ \alpha_4 = (\alpha_1 + \alpha_2 + \alpha_3)/4$
   Setting $\alpha_1 = 0$ *for identifiability we would have the matrix*

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -1/2 & 1 & 0 \\ 0 & -1/3 & -1/3 & 1 \end{bmatrix}$$

   You can also get this by coding $\mu_1 = \beta_0$, $\mu_2 = \beta_0 + \beta_1$, $\mu_3 = \beta_0 + \beta_2$ *and* $\mu_4 = \beta_0 + \beta_3$.

   (b) In fact, the levels $\alpha_1, \ldots, \alpha_4$ corresponded to measuring wool in seasons 1, 2, 3 and 4. We expect a linear increase in wool production. Find a contrast matrix to test the hypothesis that $\alpha_1 = \delta_0 + \delta_1,\ \alpha_2 = \delta_0 + 2\delta_1,\ \alpha_3 = \delta_0 + 3\delta_1,\ \alpha_4 = \delta_0 + 4\delta_1$ for some (unknown) δ_0, δ_1.
   Here the idea is to see how far $\alpha_1, \ldots, \alpha_4$ *lie from being on a straight line. Defining* $s = (1, 2, 3, 4)^T$ *and* $X = [1, s]$ *and* $H = X(X^TX)^{-1}X^T$ *we have that* $(I - H)\alpha$ *takes the residuals after performing linear regression to predict* α *from* s.
   In terms of the β, *we can specify* $\alpha_1 = 0$, $\alpha_2 = \beta_1$ *etc.*

3. In our model from Part 1,

   (a) Write down the joint distribution of $\bar{y}_{11\cdot}$ and γ_{11}, hence find the distribution of $\gamma_{11}|\bar{y}_{11\cdot}$.

$$\bar{y}_{11\cdot} = \alpha_i + \gamma_{11} + \bar{e}_{11\cdot}$$

so

$$\begin{pmatrix} \bar{y}_{11\cdot} \\ \gamma_{11} \end{pmatrix} \sim N\left(\begin{bmatrix} \mu+\alpha_i \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_g^2+\sigma_e^2/4 & \sigma_g^2 \\ \sigma_g^2 & \sigma_g^2 \end{bmatrix}\right)$$

so that

$$\gamma_{11}|\bar{y}_{11\cdot} \sim N\left(\frac{\sigma_g^2}{\sigma_g^2+\sigma_3^2/r}(\bar{y}_{11\cdot}-\alpha_i), \sigma_g^2-\frac{\sigma_g^4}{\sigma_g^2+\sigma_e^2/r}\right)$$

and below we will re-write the variance as $\sigma_g^2\sigma_e^2/r/(\sigma_g^2+\sigma_e^2/r)$.

(b) What is the distribution of $(\gamma_{11}+\gamma_{12})/2|(\bar{y}_{11\cdot}, \bar{y}_{12\cdot})$?
Since γ_{11} *is independent of* $\bar{y}_{12\cdot}$ *and vice versa, we can treat* $\gamma_{11}|\bar{y}_{11\cdot}$ *and* $\gamma_{12}|\bar{y}_{12\cdot}$ *as independent and conclude (after some algebra) that*

$$(\gamma_{11}+\gamma_{12})/2|(\bar{y}_{11\cdot}, \bar{y}_{12\cdot}) \sim N\left(\frac{\sigma_g^2}{\sigma_g^2+\sigma_3^2/r}((\bar{y}_{11\cdot}+\bar{y}_{12\cdot})/2-\alpha_i), \frac{1}{2}\sigma_g^2\sigma_e^2/(\sigma_g^2+\sigma_e^2/r)\right)$$

(c) Despite it being random, a colleague wants to provide a confidence interval for $(\gamma_{11}+\gamma_{12})/2$. What would you provide?
Using the result in b, we can create an interval in the form of

$$\frac{\sigma_g^2}{\sigma_g^2+\sigma_3^2/r}((\bar{y}_{11\cdot}+\bar{y}_{12\cdot})/2-\alpha_i) \pm 2\sqrt{\frac{1}{2}\sigma_g^2\sigma_e^2/(\sigma_g^2+\sigma_e^2/r)}$$

4. Longitudinal models

In a new experiment each sheep in three different breeds was measured in seasons 1, 2, 3 and 4. The researchers believe that each sheep increases its yield linearly, but with a different linear regression line for each sheep. Different breeds may differ in their average regression lines.

(a) Write down a model to express this understanding.

$$y_{ijk} = \alpha_i + \beta_i s_k + b_{0j} + b_{1j}s_k + e_{ijk}$$

for the k*th season* s_k *of the* j*th sheep in breed* i
Here α_i *and* β_i *are the fixed breed specific slope and intercept and* $b_{0j} \sim N(0, \sigma_{b_0}^2)$ *is a sheep-specific random intercept and* $b_{1j} \sim N(0, \sigma_{b_1}^2)$ *is a sheep-specific random slope with errors* $e_{ijk} \sim N(0, \sigma_e^2)$.

(b) Write down the covariance matrix for the responses from the jth sheep in group i:

$$\text{cov}(y_{ij\cdot}) = \sigma_{b_0}^2 J_4 + \sigma_{b_1}^2 ss^T + \sigma_e^2 I_4$$

(c) Re-write your model and the covariance with season treated as a category rather than continuous. Can you still estimate an interaction between sheep and season?

$$y_{ijk} = \alpha_i + \beta_k + \alpha\beta_{ik} + b_{0j} + \epsilon_{ijk}$$

with covariance

$$cov(y_{ij\cdot}) = \sigma^2_{b_0} J_4 + \sigma^2_e I_4$$

where we note that because each season/sheep pair only has one observation, we cannot distinguish this interaction from σ^2_e.

(d) (More abstract and more difficult). In the general longitudinal model $y = X\beta + Zb + e$, $e \sim N(0, \sigma^2 I)$, $b \sim N(0, \sigma^2 G)$

i. Write down the covariance of $X^T(I + ZGZ^T)^{-1}y$ and b.
Writing $V = (I + ZGZ^T)$ *we first expand*

$$X^T V^{-1} y = X^T V^{-1} X\beta + X^T V^{-1} Zb + X^T V^{-1} e$$

and observing that the covariance of y *is* $\sigma^2 V$ *so that* $cov(X^T V^{-1} y) = \sigma^2 X^T V^{-1} X$ *we have the joint covariance is*

$$\sigma^2 \begin{pmatrix} X^T V^{-1} X & X^T V^{-1} ZG \\ GZ^T V^{-1} X & G \end{pmatrix}$$

ii. Hence find the distribution of $b|X^T(I + ZGZ^T)^{-1}y$.
Normal with mean

$$GZ^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} (y - X\beta)$$

and covariance

$$G - GZ^T V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1} ZG$$