BTRY/STSCI 4030 - Linear Models with Matrices - Fall 2017
Final - Monday, December 11

NAME:

NETID:

Instructions:

It is not necessary to complete numerical calculations (using a calculator) if you clearly show how the answer can be obtained, and if the exact answer is not required in subsequent parts.

A set of formulae and notes is provided with the exam; other outside material is not allowed. You may directly use any result on the notes without proving it.

You may reference any result in the formulae by it's number; e.g. the Eigen-decomposition for a symmetric matrix is in 5.2a.

---

1. **Testing Variance Components**

Here we will consider the problem of testing whether a random effect really exists.

As an example, Giles is a persistent, if not very successful, grower of chilis (the Cayenne that he took indoors to try and keep it over the winter died last week, and the summer crop was pretty sad, but he'll try again next year). He's interested in how much plants versus fruit contribute to the heat of the chilis as measured by capsaicin content.

To examine this, he takes a plants and collects g fruit from each plant. He then sends r samples of each fruit to be tested for capsaicin content. This strategy lets him distinguish plant-to-plant variability from fruit-to-fruit and extract measurement error by using repeated measurements of each fruit.

This gives a model for the kth measurement of the jth fruit from the ith plant as:

$$y_{ijk} = \mu + \alpha_i + \gamma_{ij} + \epsilon_{ijk},\ i = 1, \ldots, a\ j = 1, \ldots, g,\ k = 1, \ldots, r \quad (1)$$

where α_i represents the effect of plant i, γ_{ij} is the effect of the jth fruit from the ith plant and ϵ_{ijk} is the measurement error for a particular sample.

Since Giles can't keep plants alive over winter, it makes sense for plants (and therefore fruit) to be random effects, meaning we would specify

$$\alpha_i \sim N(0, \sigma_\alpha^2),\ \gamma_{ij} \sim N(0, \sigma_g^2),\ \epsilon_{ijk} \sim N(0, \sigma_e^2)$$

Giles wonders whether there really is any fruit-to-fruit variability within a plant. That is, that all the $\gamma_{ij} = 0$. Since these are random, this is equivalent to saying $\sigma_g^2 = 0$.

Here we will try to test this hypothesis.

(a) First of all, the natural estimate of μ is $\bar{y}_{\cdots}$. What is its variance?

$$\sigma_a^2/a + \sigma_g^2/ag + \sigma_e^2/agr$$

(b) Turning to the question of interest, we would like to look at the effect of γ_{ij}. To do this we need to at least remove the α. Write down an expression for $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot}$ in terms of the right hand side of (1).

Find the expected value of

$$SSG = \sum_{i=1}^{a}\sum_{j=1}^{g}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$$

This is

$$\begin{aligned} SSG &= \sum_{i=1}^{a}\sum_{j=1}^{g}((\gamma_{ij} + \bar{\epsilon}_{ij\cdot}) - (\bar{\gamma}_{i\cdot} + \bar{\epsilon}_{i\cdot\cdot}))^2 \\ &= a(g-1)(\sigma_g^2 + \sigma_e^2/r) \end{aligned}$$

(c) We are interested in testing the hypothesis $H_0 : \sigma_g^2 = 0$. What is the distribution of $SSG/(\sigma_e^2/r)$ when H_0 is true?

Setting $z_{ij} = (\gamma_i j + \bar{\epsilon}_{ij\cdot})/(\sigma_e/\sqrt{r}) \sim N(0,1)$ *when* H_0 *is true, we have*

$$SSG/(\sigma_e^2/r) = \sum_{i=1}^{a} z_{i\cdot}^T C_g z_{i\cdot} \sim \chi^2_{a(g-1)}$$

(d) Show that the estimate

$$MSE = \frac{1}{ag(r-1)}\sum_{i=1}^{a}\sum_{j=1}^{g}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij\cdot})^2$$

is an unbiased estimate of σ_e^2.

$$\begin{aligned} MSE &= \frac{1}{ag(r-1)}\sum_{i=1}^{a}\sum_{j=1}^{g}\sum_{k=1}^{r}(\epsilon_{ijk} - \bar{\epsilon}_{ij\cdot})^2 \\ &= \frac{1}{ag(r-1)} ag(r-1)\sigma_e^2 = \sigma_e^2 \end{aligned}$$

bonus Show that $\bar{\epsilon}_{ij\cdot}$ is independent of the vector $(\epsilon_{ij1}-\bar{\epsilon}_{ij\cdot},\ldots,\epsilon_{ijr}-\bar{\epsilon}_{ij\cdot})$. Hence show that $rMSG/MSE$ has an F statistic and give its degrees of freedom and non-centrality parameter when H_0 is not true.

In this case, we have that the vector is expressed as $C_r\epsilon_{ij\cdot}$ *and* $\bar{\epsilon}_{ij\cdot} = \mathbf{1}_r^T\epsilon_{ij\cdot}$ *and these have correlation* $\sigma_e^2 C_r\mathbf{1}_r = 0$ *so are independent.*

Now we note that MSE is expressed in terms of the vector and SSG in terms of $\bar{\epsilon}_{ij\cdot}$ *so these are also indpendent and hence under* H_0, $rMSG/MSE \sim F^{a(g-1)}_{ag(r-1)}$.

Noncentrality is more tricky here. In fact, when $\sigma_g^2 \neq 0$, *we have that*

$$\frac{\sigma_e^2/r}{\sigma_g^2+\sigma_e^2/r}\frac{MSG}{MSE} \sim F^{a(g-1)}_{ag(r-1)}(0)$$

where the F *is central, but scaled up by a factor of* $(\sigma_g^2+\sigma_e^2/r)/(\sigma_e^2/r)$. *Note that we were very liberal in grading this bit.*

2. **Contrasts for Linear Trend**

A lot of Giles' problems are associated with Ithaca, where the summer weather can change a lot from one year to the next. Nonetheless, he hopes that by continuing to fertilize his garden, the soil is getting better for the plants over time.

To assess this, he collects the total weight of chilis produced in each of five plants each year.

Because there is so much change from year to year (both up and down), he has coded year as a categorical effect with one level for each of $2013, \ldots, 2017$, which have levels $\alpha_1, \ldots, \alpha_5$. But he's interested in whether the total weight tends to increase over time, despite a lot of year-to-year variation. By centering the years he gets the vector $\boldsymbol{s} = (-2, -1, 0, 1, 2)$ to indicate year by a continuous value.

(a) Use $\boldsymbol{s}$ to write down a contrast of the α's to test for whether the there is a long-term increase in yield over time, even if not lying exactly on a straight line.

Why is this coding of $\boldsymbol{s}$ useful instead of, say, (2013,2014,2015,2016,2017)?

$L = \boldsymbol{s}^T$ – this measures the correlation between $\boldsymbol{\alpha}$ and time. It's equivalent to looking at the slope when $\boldsymbol{\alpha}$ is regressed on $\boldsymbol{s}$, but since $\boldsymbol{s}$ is centered, we already remove a constant level over time.

(b) Since this model is to be fit in `R`, the effects will actually be reported in reference coding using the first level (year 2013) as a reference. What is the equivalent contrast that you would apply here?

Refitting in effect coding gives $L = (0, -1, 0, 1, 2)$.

(c) The contrast above is not different from zero at the 0.05 level. This leads you to conclude:

i. That all the levels are the same.
ii. That there is no linear trend in the years.
iii. That all the levels line up on a straight line in $\boldsymbol{s}$.
iv. Not much of anything.

Choose one response.

Correct answer is not much of anything – you can't accept the null hypothesis (which would be (ii)), we aren't looking at the residuals from a straight line (iii) and (i) is even stronger than (i).

bonus It's been suggested that since Giles can't really repeat 2016 all over again (although there are several reasons he'd like to), year really should be treated as a random effect. What analysis would you carry out in this case?

An appropriate model would be

$$y_{ij} = \beta_0 + \beta_1 s_j + b_j + \epsilon_{ij},\ b_j \sim N(0, \sigma_b^2),\ \epsilon_{ij} \sim N(0, \sigma_e^2)$$

that is we treat year as linear in the fixed effect equation and categorical for the random component.

3. **Diagnostics and Hypothesis Tests**

One standard diagnostic is the deletion residual, $e_{(i)} = y_i - \boldsymbol{x}_i^T \hat{\beta}_{(i)}$. That is the error from a prediction that did not get to use y_i. The studentized version of this is

$$r_i = \frac{y_i - \boldsymbol{x}_i^T \hat{\beta}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(1 + \boldsymbol{x}_i (X_{-i}^T X_{-i})^{-1} \boldsymbol{x}_i)}} \sim t_{n-p-2}$$

A suggestion is to conduct a hypothesis test of whether each r_i is larger than it should be. That is, to remove any observation i for which $r_i > t^{0.05}_{n-p-2}$. This would

(a) Identify observations that are further away from the regression line than you would expect by chance.

(b) Indicate that there is curvature in your model.

(c) Suggest which coefficients are unduly influenced by which observations.

(d) Throw out about 5% of the observations, even if all regression assumptions were met.

Choose one response.

Throw out 5% of observations: that's the definition of a hypothesis at the 5% level. Partial credit for (a), but we do expect 5% of observations to be larger than the critical value.

4. **Two-Stage Longitudinal Models**

Here we will consider estimating a longitudinal model in a two-step manner. First, we estimate a linear regression to get a slope and intercept for each subject. Then we use the estimated slopes and intercepts as data.

In this case, Giles is still obsessed with not-very-successful chili plants. These are normally planted in June and grow through October. He measures the height of each of a plants on the first of each month, starting in June, to give five measurements per plant.

We assume that the plants grow linearly over time, and each plant has its own (random) slope and intercept. We'll code month as $\boldsymbol{m} = (-2, -1, 0, 1, 2)^T$ and let $X = [\mathbf{1}, \boldsymbol{m}]$. This results in a model for the five measurements $\boldsymbol{y}_i$ of plant i as

$$y_{ij} = \beta_0 + \beta_1 m_j + b_{0i} + b_{1i} m_j + \epsilon_{ij}$$

or

$$\boldsymbol{y}_i = X\boldsymbol{\beta} + X\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,\ \boldsymbol{b}_i \sim N(\mathbf{0}, G),\ \boldsymbol{\epsilon}_i \sim N(0, \sigma^2 I).$$

(a) We consider estimating a line for each plant separately from

$$\hat{\boldsymbol{\beta}}_i = \begin{pmatrix} \hat{\beta}_{0i} \\ \hat{\beta}_{1i} \end{pmatrix} = (X^T X)^{-1} X^T \boldsymbol{y}_i$$

What is the mean and variance of $\hat{\boldsymbol{\beta}}_i$?
Writing

$$\begin{aligned}(X^T X)^{-1} X^T \boldsymbol{y}_i &= (X^T X)^{-1} X^T X\boldsymbol{\beta} + (X^T X)^{-1} X^T X\boldsymbol{b}_i + (X^T X)^{-1} X^T X\boldsymbol{\epsilon}_i \\ &= \boldsymbol{\beta} + \boldsymbol{b}_i + (X^T X)^{-1} X^T X\boldsymbol{\epsilon}_i\end{aligned}$$

from which

$$E\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta},\ var(\hat{\boldsymbol{\beta}}_i) = G + \sigma^2 (X^T X)^{-1}$$

(b) Write down the covariance of $\hat{\boldsymbol{\beta}}_i$ and $\boldsymbol{b}_i$ and hence calculated the conditional expectation $E(\boldsymbol{b}_i | \hat{\boldsymbol{\beta}}_i)$.

Using the results above

$$cov(\hat{\boldsymbol{\beta}}_i|\boldsymbol{b}_i) = cov(\boldsymbol{b}_i, \boldsymbol{b}_i) = G$$

and therefore

$$E(\boldsymbol{b}_i, \hat{\boldsymbol{\beta}}_i) = G(G + \sigma^2(X^TX)^{-1})^{-1}(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})$$

(c) To get at the pure error term, we consider the residuals from each plant-specific regression. That is

$$\boldsymbol{e}_i = (I - X(X^TX)^{-1}X^T)\boldsymbol{y}_i$$

What are the mean and variance of the $\boldsymbol{e}_i$?
Observing that

$$(I - X(X^TX)^{-1}X^T)\boldsymbol{y}_i = (I - X(X^TX)^{-1}X^T)\boldsymbol{\epsilon}_i$$

because $(I - H)X = 0$*, we have*

$E(I-X(X^TX)^{-1}X^T)\boldsymbol{y}_i = 0,\ var((I-X(X^TX)^{-1}X^T)\boldsymbol{y}_i) = \sigma_e^2(I-H)$

because $(I - H)$ *is idempotent.*

(d) We can estimate the mean and variance of $\hat{\boldsymbol{\beta}}_i$ by using them as data. This gives us unbiassed estimates

$$\bar{\boldsymbol{\beta}} = \frac{1}{a}\sum_{i=1}^{a}\hat{\boldsymbol{\beta}}_i,\ \hat{V}_\beta = \frac{1}{a-1}\sum_{i=1}^{a}(\hat{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}})^T.$$

for mean and variance respectively.

Suggest an unbiassed estimate of σ^2, hence give an estimate of $G = \text{var}(\boldsymbol{b}_i)$.

Since $\boldsymbol{e}_i = (I - H)\boldsymbol{\epsilon}_i$*, we have an unbiassed estimate*

$$\hat{\sigma}_e^2 = \frac{1}{3a}\sum_{i=1}^{a}\boldsymbol{e}_i^T\boldsymbol{e}_i$$

where $3 = 5 - 2 = tr(I - H)$ *because there are 5 observations per plant.*

With this, an unbiased estimate of G *is given by*

$$\hat{G} = V_\beta - \hat{\sigma}_e^2(X^TX)^{-1}$$

Since V_β *is unbiased for* $G + \sigma^2(X^TX)^{-1}$.

bonus These calculations give the ReML estimates when the data is balanced. However, when the observation times are not the same for each plant, this becomes more difficult. That is there is a different matrix X_i for each plant. What is the covariance of

$$\hat{\boldsymbol{\beta}} = \frac{1}{a}\sum_{i=1}^{a}\hat{\boldsymbol{\beta}}$$

in this case? How would you insist that b_{0i} and b_{1i} are uncorrelated? Or drop b_{1i} and just have a random intercept?

$$\mathrm{var}(\bar{\boldsymbol{\beta}}) = \frac{1}{a^2}\sum_{i=1}^{a}\mathrm{var}(\hat{\boldsymbol{\beta}}_i) = G/a + \frac{\sigma^2}{a}\sum_{i=1}^{a}(X_i^T X_i)^{-1}$$

Note that the off-diagonals for V_β will not necessarely be zero, which means we would estimate there to be a correlation between b_{0i} and b_{1i}. To obtain an estimate with them independentent, we could simply set the diagonals to zero in our estimates.

To drop the random slope; we could use $\bar{\boldsymbol{\beta}}$ as a fixed slope and intercept, and then produce a constant plant effect by taking the average residual for each plant.