York University

MATH 6642 – Midterm Test

Instructor: Georges Monette

May 21, 2019 – 5:45 pm to 6:45 pm (60 minutes)

WARNING

DO NOT OPEN THIS BOOKLET
UNTIL YOU ARE
INSTRUCTED TO DO SO

Student number: ____

Family name: (in BLOCK letters)____

Given name: (in BLOCK letters)____

Signature____

Information:

Be sure to read questions closely. Some may ask for multiple pieces of information. Make sure to respond completely. If you need more space to answer, write "**OVER**" and continue the answer on the back of the page.

The marks for each questions are shown at the end of the question. The sum of the marks is 110. The exam will be graded out of 100 so that you can potentially earn 10 bonus points.

Aids allowed: Non-programmable calculator, ruler, pencils, pens, erasers.

WARNING

DO NOT OPEN THIS BOOKLET
UNTIL YOU ARE
INSTRUCTED TO DO SO

1. Prove that the following are equivalent:
   a) A is a variance matrix
   b) A is non-negative definite
   c) all the eigenvalues in the spectral decomposition of A are non-negative
   d) there exist a matrix B such that $A = BB'$
   Note: you may use the spectral decomposition theorem. *(25 points)*

Let A be $n \times n$

(a) $\Rightarrow$ (b)

$\exists$ random vector $\underset{\sim}{X} \ni \mathrm{Var}(\underset{\sim}{X}) = A$

Let $\underset{\sim}{c} \in \mathbb{R}^n$. Then

$$0 \leq \mathrm{Var}(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}' A \underset{\sim}{c}$$

$\therefore$ A is NND

(b) $\Rightarrow$ (c) Using the SpDT $A = \Gamma \Lambda \Gamma'$

Let γ_i be a column of Γ.

Then $\gamma_i' \Gamma \Lambda \Gamma' \gamma_i = \lambda_i$

which is ≥ 0 since A is NND.

(c) $\Rightarrow$ (d)

Since $\lambda_i \geq 0$ for all i we can express

$$\Lambda = \Lambda^{1/2} \Lambda^{1/2}$$

Let $B = \Gamma \Lambda^{1/2}$

then $BB^T = \Gamma \Lambda \Gamma' = A$

(d) $\Rightarrow$ (a) Let $\underset{\sim}{X}$ be normal $N_n(\underset{\sim}{0}, I)$

Let $\underset{\sim}{Y} = B\underset{\sim}{X}$. Then $\mathrm{Var}(\underset{\sim}{Y}) = BIB' = A$.

2. The following output shows an analysis predicting math achievement (mathach) using socio-economic status (ses) and sex (Male or Female) in a sample of high school students from a number of public schools.

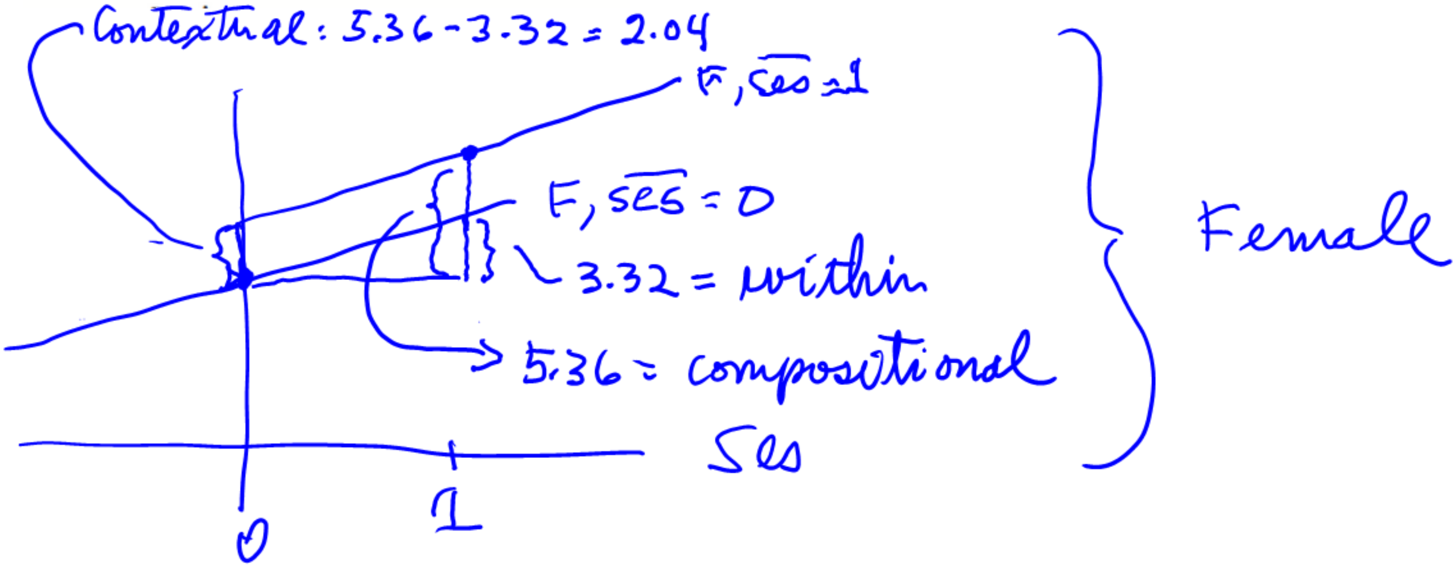
```
fit <- lme(mathach ~ (dvar(ses, school) + cvar(ses, school)) * Sex,
          subset(hs, Sector == 'Public'),
          random = ~ 1 + dvar(ses, school) | school)
summary(fit)
```

```
Linear mixed-effects model fit by REML
 Data: subset(hs, Sector == "Public")
       AIC      BIC    logLik
  5481.758 5528.936 -2730.879

Random effects:
 Formula: ~1 + dvar(ses, school) | school
 Structure: General positive-definite, Log-Cholesky parametrization
                  StdDev    Corr
(Intercept)       1.1678502 (Intr)
dvar(ses, school) 0.9334333 0.631
Residual          6.3754705

Fixed effects: mathach ~ (dvar(ses, school) + cvar(ses, school)) * Sex
                              Value Std.Error  DF   t-value p-value
(Intercept)               11.736739 0.4359394 810 26.922868  0.0000
dvar(ses, school)          3.324965 0.5433705 810  6.119150  0.0000
cvar(ses, school)          5.364838 0.9201634  17  5.830310  0.0000
SexMale                    1.245271 0.4710892 810  2.643387  0.0084
dvar(ses, school):SexMale -0.900169 0.6953643 810 -1.294529  0.1959
cvar(ses, school):SexMale  0.978971 1.0631268 810  0.920841  0.3574
 Correlation:
                          (Intr) dv(,s) cv(,s) SexMal d(,s):
dvar(ses, school)          0.196
cvar(ses, school)          0.326 -0.004
SexMale                   -0.538 -0.039 -0.188
dvar(ses, school):SexMale -0.029 -0.660  0.016  0.007
cvar(ses, school):SexMale -0.172  0.012 -0.523  0.317  0.003

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.59491506 -0.78060374  0.04117356  0.75209401  2.65893697

Number of Observations: 833
Number of Groups: 19
```

Draw a graph with four lines representing the expected value of math achievement as a function of individual ses under the following conditions:
1) female students in a school with mean ses equal to 0,
2) male students in a school with mean ses equal to 0,
3) female students in a school with mean ses equal to 1,
4) male students in a school with mean ses equal to 1.
On the graph, show clearly the value and position of the contextual, compositional, and within-school effects of 'ses' for both male and female students. Use the back of the previous page to draw the graph. *(25 points)*

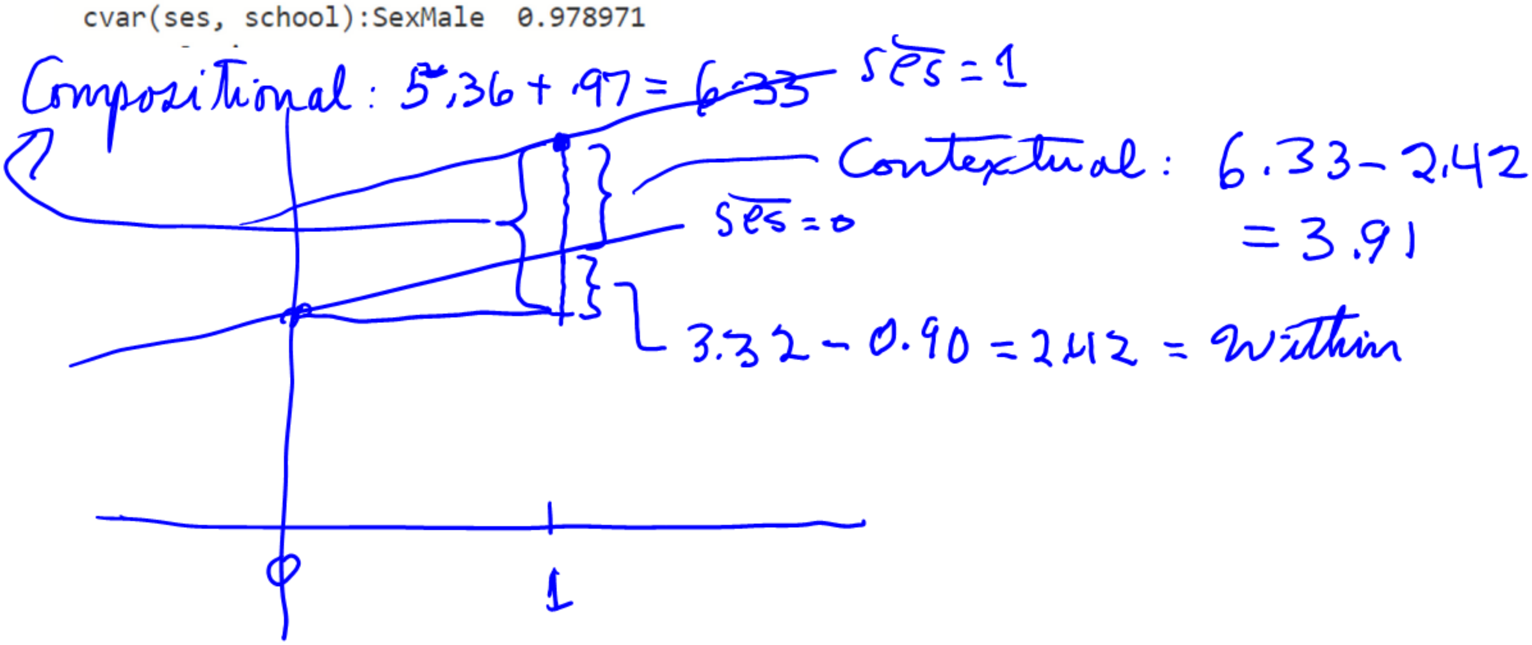
```
Fixed effects: mathach ~ (dvar(ses,
                              Value
(Intercept)               11.736739
dvar(ses, school)          3.324965
cvar(ses, school)          5.364838
SexMale                    1.245271
dvar(ses, school):SexMale -0.900169
cvar(ses, school):SexMale  0.978971
```

Female




```
Fixed effects: mathach ~ (dvar(ses,
                              Value
(Intercept)               11.736739
dvar(ses, school)          3.324965
cvar(ses, school)          5.364838
SexMale                    1.245271
dvar(ses, school):SexMale -0.900169
cvar(ses, school):SexMale  0.978971
```

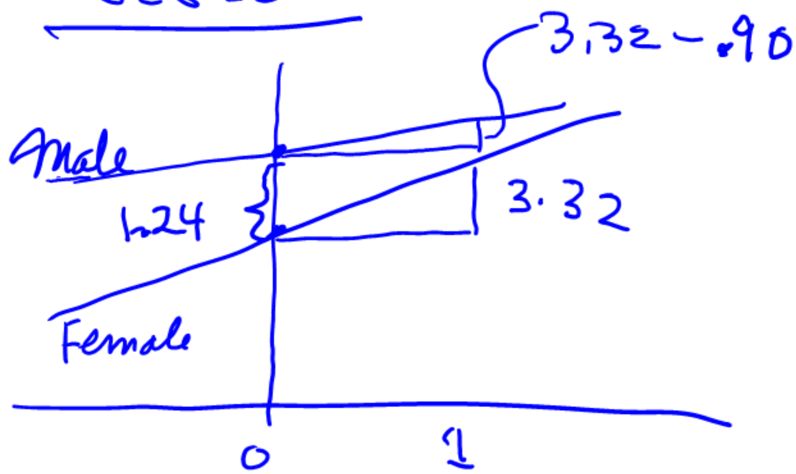
Male



3. (continued from the previous question) Draw a suitable graph to show clearly the value and position of the 'gender gap' for students with ses equal to 0 in a school with mean ses equal to 0, for students with ses equal to 1 in a school with mean ses equal to 0, and for students with ses equal to 1 in a school with mean ses equal to 1. *(20 points)*

```
Fixed effects: mathach ~ (dvar(ses,
                              Value
(Intercept)                11.736739
dvar(ses, school)           3.324965
cvar(ses, school)           5.364838
SexMale                     1.245271
dvar(ses, school):SexMale -0.900169
cvar(ses, school):SexMale  0.978971
```

$\overline{ses}=0$

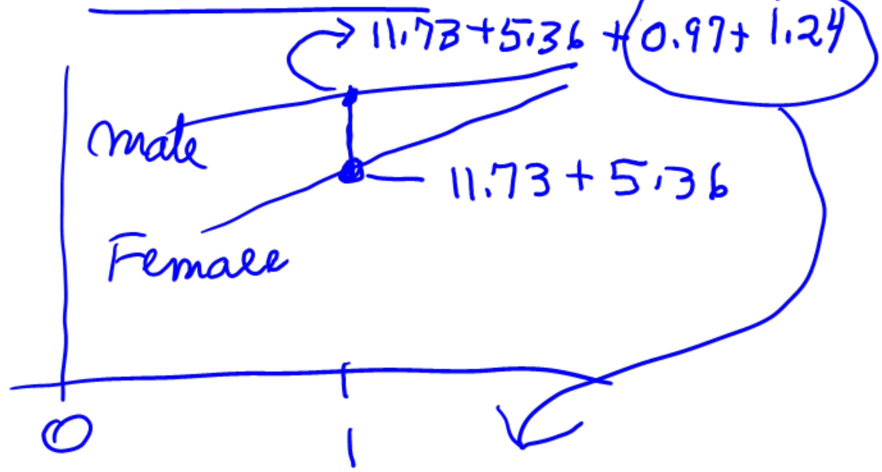


$\overline{ses}=0$

Gap at ses = 0 : 1.24

Gap at ses = 1 : 1.24 − 0.9 = 0.34

$\overline{ses}=1$



$\overline{ses}=1$

Gap at ses = 1 : 0.97 + 1.24 = 2.21

(∵ $\overline{ses}=1$

so dvar = ses − $\overline{ses}$ = 0)

4. Consider a linear regression of a continuous numerical variable Y on four continuous numerical predictors: X_1, X_2, X_3, X_4 and an intercept term. Let X_h be the predicted value of X_1 using a linear regression of X_1 on X_2, X_3, X_4, and let Y_h be the predicted value of Y using a linear regression of Y on X_2, X_3, X_4 and an intercept. Let $X_r = X_1 - X_h$, $Y_r = Y - Y_h$. To answer the following you may use, in addition to basic linear algebra, any of the 'three basic theorems' and any result presented in the 'Outline of Linear Algebra for Regression'.

a) Show that the simple regression coefficient of Y on X_r is equal to the regression coefficient for X_1 in the regression of Y on X_1, X_2, X_3, X_4 and an intercept term.

b) Show which model results in a smaller SSE: (1) `Yr ~ Xr` or (2) `Y ~ X + Xh` *(20 points)*

a) By the AVP theorem, the reg. coeff of Y on X_1 in the model $Y \sim X_1 + X_2 + X_3 + X_4$ is the same as the reg. coeff of Y_r on X_r with the same SSE.

So $Y_r = X_r \hat{\beta}_1 + e \qquad e \perp X_r = QX_1$

where Q projects onto $\mathcal{L}^{\perp}(1, X_2, X_3, X_4)$

Now $Y = Y_r + (Y - Y_r)$

$= X_r \hat{\beta}_1 + e + (Y - Y_r)$

$= X_r \hat{\beta}_1 + e + PY$

where P projects on $\mathcal{L}(1, X_2, X_3, X_4)$

and $Y'PX_r = Y'PQX_1 = Y'0X_1 = 0$

So $Y = X_r \hat{\beta}_1 + d$ where $d \perp \mathcal{L}(X_r)$

and thus $\hat{\beta}_1$ is the reg. coeff of Y on X_r

(b) From the AVP and linear propensity score theorems we know that they both have the same reg. coeff. as the full model $Y \sim X_1 + X_2 + X_3 + X_4$

So $Y_p = X_p \hat{\beta}_1 + e$ (1)

and $Y = X_1 \hat{\beta}_1 + X_n \hat{\beta}_2 + d$ (2)

The AVP for (2) has the values of X_p as (1) since $X_p = X_1 - X_n$ is the residual of X_1 on X_n

So $Y_* = X_p \hat{\beta}_1 + f$ $\quad f \perp \mathcal{L}(X_r)$

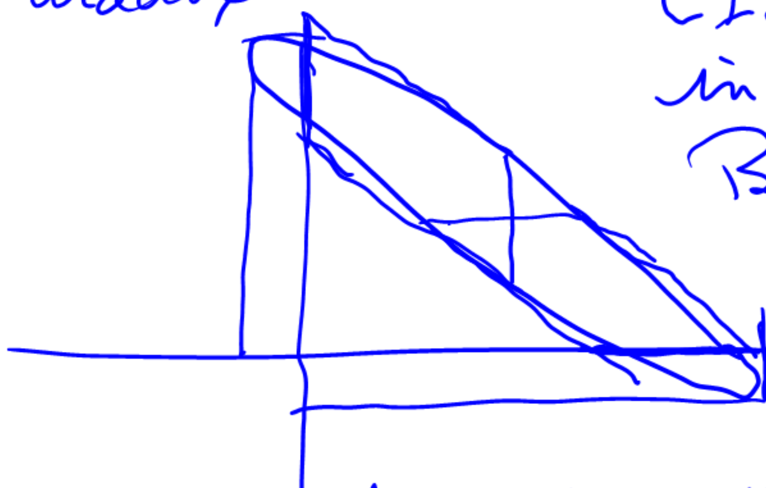
where Y_* is the residual from regressing Y on $X_n \in \mathcal{L}(X_2, X_3, X_4)$.

Thus SSE for Y_* on X_p is at least as large as SSE for Y_p on X_p which is equal to the SSE of Y on $X_1, \ldots, X_4$.

Thus the SSE of model (1) is at least as small as the SSE of model (2).

5. Consider a linear regression of a continuous numerical variable Y on two continuous numerical predictors: X_1, X_2. Discuss whether it is possible for the multiple regression coefficients of Y on X_1, X_2 to be non-significant although the simple regression coefficients of Y on X_1 and of Y on X_2 are highly significant. Use the geometry of the confidence ellipse for β_1, β_2 and the confidence intervals for the simple regression coefficients to justify your answer. *(20 points)*

This is a example of the Dylan Paradox



CIs for β_1 & β_2 include 0

But the CIs for simple reg. coeffs exclude 0

Note that dropping either X_1 or X_2, although neither is significant does have an impact on the estimated coefficient and the significance of the other variable although the new estimated coefficient remains within the CI of the original model.