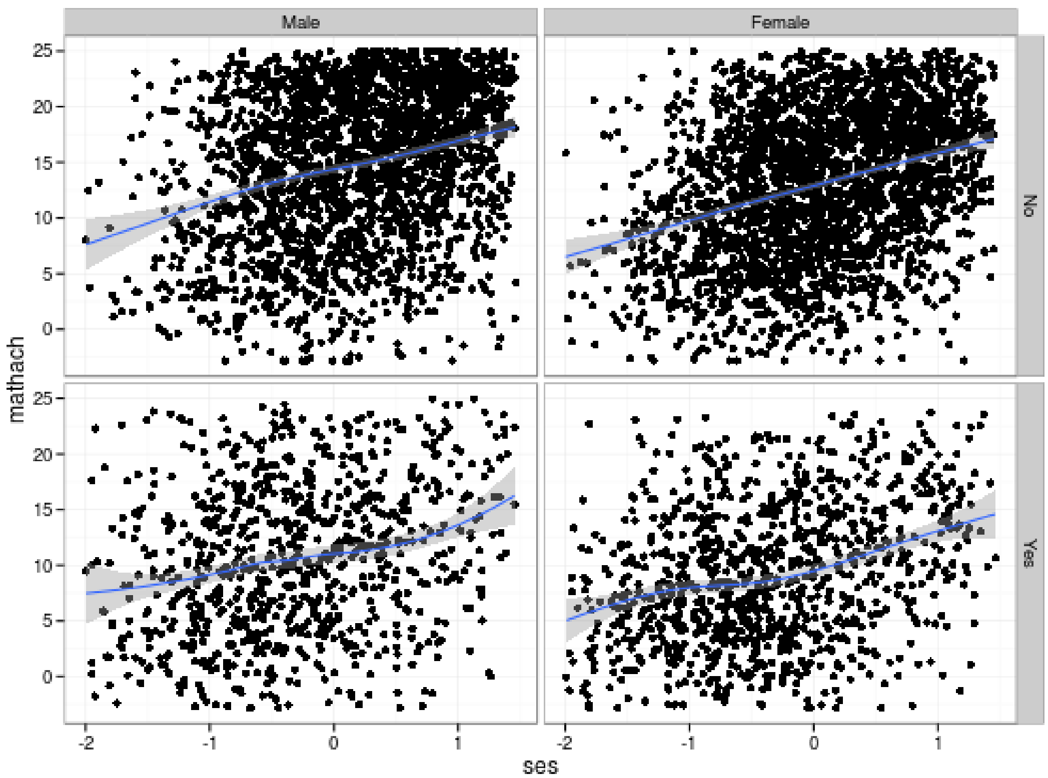
STAT 506 Midterm
March X, 201Y

50 points Name ______________________

1. A math achievement test was given to 7054 high school students in 160 different schools. The score is our response, $y$. As predictor variables, we have: $x_1$: SES (socio-economic status, a numeric variable from -2 to 1.5), $x_2$ gender, and $x_3$ minority status (yes or no), all three being measured for each individual student.

```
qplot(x=ses, y = mathach, data = myMathAch, geom=c("point","smooth")) +
      theme_bw() + facet_grid(minority ~ sex)
```




Use this model and the associated `lmer` output.

$$\boldsymbol{y}_j = [\mathbf{1}\ \boldsymbol{x}_{1j}\ \boldsymbol{x}_{2j}\ \boldsymbol{x}_{3j}] \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + [\mathbf{1}\ \boldsymbol{x}_{1j}] \begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} + \boldsymbol{\epsilon}_j,\ j = 1, \ldots, 160$$

```
ma.fit1 <- lmer(mathach ~ ses + sex + minority + (1+ ses | school), myMathAch)
 print(ma.fit1, correlation=FALSE)


   Linear mixed model fit by REML ['lmerMod']
   Formula: mathach ~ ses + sex + minority + (1 + ses | school)
      Data: myMathAch
   REML criterion at convergence: 45577
```

```
Random effects:
 Groups   Name        Std.Dev. Corr
 school   (Intercept) 1.914
          ses         0.574    -0.41
 Residual             5.994
Number of obs: 7054, groups: school, 160
Fixed Effects:
(Intercept)          ses    sexFemale  minorityYes
      14.16         2.11        -1.20        -2.99
```

(a) Interpret estimated population effects for ses, minority, and sex (skip the intercept). Give the estimated size, sign, and the strength of evidence for each being nonzero. Explain in terms of this situation, not just in general platitudes. (10pts)

(b) Write out the estimated distribution of $\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix}$. Plug in the estimates from the output. (5 pts)

(c) Ses appears in the model twice. Explain the difference in interpretation between the two ses outputs. (4 pts)

(d) Let $\boldsymbol{y}_1$ be the vector of 15 math achievement scores for the first school (labeled 8367 below).

```
xtable(head(ranef(ma.fit1)[["school"]]))
```

| | (Intercept) | ses |
|---|---|---|
| 8367 | -4.98 | 0.46 |
| 8854 | -4.99 | 0.65 |
| 4458 | -2.38 | 0.23 |
| 5762 | -2.83 | 0.36 |
| 6990 | -2.84 | 0.14 |
| 5815 | -2.42 | 0.37 |

i. Write down its predicted mean vector (based on $\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j}, \boldsymbol{x}_{3j}$ which you don't have access to for this exam, just use them as symbols) plugging in the estimates from the output. (5 pts)

ii. What is the estimated variance-covariance of $\boldsymbol{y}_1$? (Hint: avoid rewriting things you already explained above by giving them a label above and just using that here.) (5 pts)

(e) If we did the same analysis in SAS PROC MIXED:

i. How would the fixed effects SAS reports differ from those of R? (3 pts)

ii. How would the random effects SAS reports differ from those of R? (3 pts)

(f) Two other models were considered. Explain how they differ and which you prefer based on the following output. (6 pts)

```
ma.fit2 <- lmer(mathach ~ ses * minority + sex + (1+ ses | school), myMathAch)
ma.fit3 <- lmer(mathach ~ ses * minority + ses * sex + (1+ ses | school), myMat
xtable(AIC(ma.fit1,ma.fit2,ma.fit3))
```

| | df | AIC |
|---|---|---|
| ma.fit1 | 8.00 | 45593.02 |
| ma.fit2 | 9.00 | 45580.31 |
| ma.fit3 | 10.00 | 45579.78 |

2. Consider a comparison of two curricula (A and B) for teaching intro stats to college students. Students will register for a section of the class which fits their schedule not knowing if the section will be taught using the A or the B curriculum.

   (a) How will this design allow for causal inference? Explain any necessary assumptions. (5 pts)

   (b) Curriculum B requires two 75 minute meeting times each week, but A will be taught in three 50 minute sessions. Discuss in terms of imbalance and/or overlap. What further assumption(s) is(are) required? (4 pts)