36-402/608 Homework #11 due 10:30AM 4/8

1. Dyads (60 points)

   You must use SAS for this problem! Use the DDFM=SATTERTH option.

   This problem is a study of income in married couples in Massachusetts. Use this code to load the data in dyads.dat.

```
DATA dyads;
    INFILE "dyads.dat" FIRSTOBS=2;
    INPUT agentNum dyadNum Agender Acollege Aage Pgender Page Pcollege income;
    Aage30 = Aage-30;
    Page30 = Page-30;
RUN;
```

   Use only the adjusted age variables to obtain a more interpretable intercept.

   The explanatory variables are gender (1=female, -1=male), college (indicator variable of college graduate), and age (in years). "A" indicates "agent" and "P" indicates partner. First do some EDA to get familiar with the study variables. Then select an appropriate dyadic random effect model. Justify the need for a random per-dyad intercept. Don't try and random slopes. Start with the largest possible fixed effects model without interactions. Remove unneeded fixed effects.

   Turn in a brief description of what you learned about the study from the EDA, the BIC values you used in your model selection, the output and SAS code for the final model including a residual plot and your interpretation of it, and a brief interpretation of the estimated parameters. Also state any one interaction that would be worth studying, and what a small p-value would indicate for that interaction.

   I see that ages average 40.2 for agents and partners, with a range of 22 to 60. One key observation is shown here:

```
PROC FREQ;
    TABLES Agender*Pgender Acollege*Pcollege dyadNum;
RUN;
```

```
Agender     Pgender

Frequency|
Row Pct  |
Col Pct  |      -1|       1|  Total
---------+--------+--------+
      -1 |     40 |      0 |     40
         | 100.00 |   0.00 |
         | 100.00 |   0.00 |
---------+--------+--------+
       1 |      0 |     40 |     40
         |   0.00 | 100.00 |
         |   0.00 | 100.00 |
---------+--------+--------+
Total          40       40       80
            50.00    50.00   100.00
```

Therefore this is a study of homosexual couples only, 40 couples for each gender.

I started with:

```
TITLE2 "No random effect";
PROC MIXED;
    CLASS Acollege Pcollege Agender;
    MODEL income = Agender Acollege Aage30 Page30 Pcollege / DDFM=SATTERTH;
RUN;
```

and found BIC=567.4 (REML). Adding the random intercept with "subjects" equal to the schools (upper level of the hierarchy), I find that the BIC drops to 565.1, so the random intercept is needed.

Use ML, I choose fixed effect as follows:

Fixed effects	BIC
Agen+Acol+Aage30+Page30+Pcol	595.2
Agen+Acol+Aage30+Pcol	591.5
Acol+Aage30+Pcol	588.1

Now all of the p-values are <0.0001, so I don't try to drop any more.

The final model is:

```
TITLE2 "REML: RI+Agender Acollege Aage30 Pcollege";
ODS GRAPHICS ON / IMAGENAME="HW11P1Res" IMAGEFMT = PDF;
PROC MIXED METHOD=REML PLOTS=STUDENTPANEL(CONDITIONAL);
```

```
    CLASS dyadNum Acollege Pcollege;
    MODEL income = Acollege Aage30 Pcollege / DDFM=SATTERTH SOLUTION;
    RANDOM INT / SUBJECT=dyadNUM VCORR;
RUN;
ODS GRAPHICS OFF;
```

Model Information

Dependent Variable	income
Subject Effect	dyadNum
Estimation Method	REML
Degrees of Freedom Method	Satterthwaite

Dimensions

Covariance Parameters	2
Columns in X	6
Columns in Z Per Subject	1
Subjects	40
Max Obs Per Subject	2
Number of Observations Used	80

Convergence criteria met.

Estimated V Correlation
Matrix for dyadNum 1

Row	Col1	Col2
1	1.0000	0.3639
2	0.3639	1.0000

Covariance Parameter Estimates

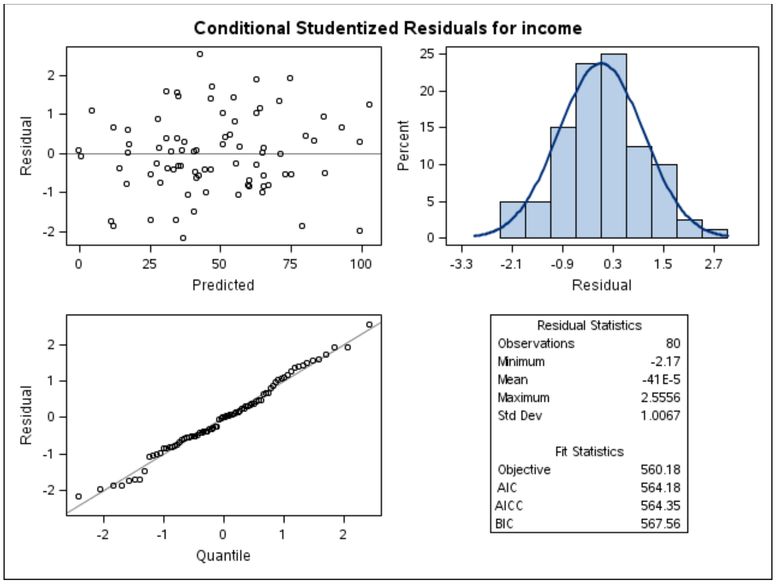
Cov Parm	Subject	Estimate
Intercept	dyadNum	28.4454
Residual		49.7215

BIC (smaller is better) 567.6

Solution for Fixed Effects

Effect	Acoll	Pcoll	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept			31.1853	2.2323	38.8	13.97	<.0001
Acollege	0		-20.5771	1.9675	67.9	-10.46	<.0001
Acollege	1		0	.	.	.	.
Aage30			2.0459	0.1134	46.3	18.04	<.0001

Pcollege	0	11.6166	1.9710	67.8	5.89	<.0001
Pcollege	1	0	.	.	.	.



The residual plot shows no problems with violations of the linearity, equal variance, or normality assumptions.

The estimated mean income averaged across gender, for couples with two 30 years olds who DID graduate from college is \$31,185. Gender and age of the partner did not have statistically significant effects on the income. There is a statistically significant association on a person's age with their income such that each additional year of age corresponds to \$2,046 more income (95%CI=[2.04+/-2(0.11)]). If a person graduated from college their income rises by \$20,600 on average, and a college graduate partner is associated with an \$11,600 decrease. The dyad-to-dyad variation in average income has standard deviation \$5,333 (square root of 28.4454). After correcting for the effects in the model, the correlation of incomes for two members of the same couple is 0.364. The residual s.d. is \$7,050.

One possible interaction is gender with Acollege. This would indicate that the benefit to a person of graduating from college differs in magnitude by gender. (Several other interactions are also worth checking.)

2. Schools (40 points)

You must use SAS for this problem! Use the DDFM=SATTERTH option.

File fundses.dat has data from a study of the relationship between a test score and school socio-economic status and funding level. Schools were randomly selected, and one randomly selected 7th grade classroom was tested from each school. The variables are school id number, school funding level, school average SES, and individual test score in that order with one line per student. The funding is in thousands of dollars per student. Adjust the funding variable so that the intercept will correspond to $10,000 per student. The per school SES variable is a z-score (mean 0, sd=1).

Run the mixed model with a random per-school intercept and fixed effects for the other explanatory variables without interaction and using REML. (You don't need to do EDA. Don't do model selection. You don't need to make residual plots.) Turn in the SAS code, the SAS output, and your interpretations of the parameters.

```
OPTIONS LINESIZE=70;
DATA fundses;
    INFILE "fundses.dat" FIRSTOBS=2;
    INPUT school funding SES score;
    fund10 = funding-10;
RUN;

PROC MIXED COVTEST;
    CLASS school;
    MODEL score = fund10 SES / DDFM=SATTERTH SOLUTION;
    RANDOM INT / SUBJECT=school;
RUN;
```

```
                          Model Information
        Dependent Variable           score
        Subject Effect               school
        Estimation Method            REML
        Degrees of Freedom Method    Satterthwaite

                             Dimensions
                 Covariance Parameters            2
                 Columns in X                     3
                 Columns in Z Per Subject         1
                 Subjects                        50
                 Max Obs Per Subject             40
                 Number of Observations Used   1465

                      Convergence criteria met.

                   Covariance Parameter Estimates
                                          Standard         Z
 Cov Parm      Subject    Estimate         Error       Value        Pr > Z
 Intercept     school       215.75       44.7876        4.82        <.0001
 Residual                  44.7136        1.6810       26.60        <.0001
```

```
             BIC (smaller is better)        9968.7

                   Solution for Fixed Effects
                         Standard
Effect       Estimate       Error      DF    t Value    Pr > |t|
Intercept     49.1969      2.1170    47.1      23.24      <.0001
fund10         8.0278      0.8268    47.1       9.71      <.0001
SES            5.7128      1.9434    47.1       2.94      0.0051
```

In this study of 1465 students in 50 schools, we find that for schools with average SES and \$10,000 per student funding, the average test score is 49.2 (SE=2.12, df=47.1, p<0.0001). Increasing funding by \$1000 per student is associated with a mean rise in test score of 8.03 (SE=0.83, df=47.1, p<0.0001). A 1 s.d. rise in school average SES is associated with a 5.72 point rise in test score (SE=1.94, df=47.1, p<0.0001). The school-to-school variation in test score for schools with the same funding and SES level has an estimated variance 216 points (s.d.=14.7), which is a substantial amount relative to the fixed intercept. The residual student-to-student variance after corrected for funding, SES, and unmeasured school variables is 44.7 (s.d.=6.7 points).