Sociology 63993
Exam 3
May 9, 2012

I. True-False. (20 points) Indicate whether the following statements are true or false. If false, briefly explain why.

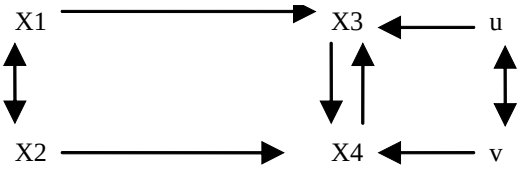
1. A researcher wants to run a logistic regression. She is using survey data in which both clustering and stratification were used in the data collection, and in which cases have differing probabilities of selection. She should therefore use likelihood ratio chi-square tests to compare and contrast models.

2. A researcher is comparing two different populations. She notes that, when Y is regressed on X, the R^2 value is larger in population 2 than it is in population 1. This could be because the structural effect of X on Y is larger in population 2 than it is in population 1.

3. The log odds of an event occurring are less than 0. That means that it is impossible for the event to occur.

4. A key advantage of fixed effects models over random effects models is that fixed effects models tend to have smaller standard errors.

5. A researcher is interested in the following model:



This model implies that X1 is uncorrelated with X4.

II. Short answer. (25 pts each, 50 pts total). Answer *both* of the following.

II-1. (25 points): As Notre Dame noted in a press release on May 7, 2012, "Last week, the Vatican charged the Leadership Conference of Women Religious (LCWR), an organization that represents most of America's Catholic nuns, with "serious doctrinal problems" and announced plans to place LCWR into a sort of receivership overseen by three American bishops." Cathleen Sprows Cummings, a professor at Notre Dame, said that "Considering the many problems facing the American church, especially the legal, moral and financial consequences of a devastating clergy sex-abuse crisis, it does seem curious that the Vatican leaders would single out Women Religious as a group in need of reform."

Several of the largest dioceses in the country are concerned about how the Vatican's actions will affect parishioners' willingness to donate to their upcoming fundraising drives. A random sample of 9000 Catholic churchgoers has therefore been interviewed. The measures include

Variable	*Description*
donate	Plans to donate to the drive (1 = yes, 0 = no)
male	Coded 1 if male, 0 otherwise
nuntaught	Coded 1 if the respondent was taught by Nuns while in school, 0 otherwise
nuns	Scale that measures how favorably respondents feel about the contributions nuns make to the Catholic Church. The higher the score, the more favorable the opinion is. The scale has been centered to have a mean of 0.

The study obtains the following results (parts of the output have been deleted):

```
. nestreg, lr: logit donate male nuntaught nuns , nolog
```

Block 1: male

```
Logistic regression                               Number of obs   =       8999
                                                  LR chi2(1)      =        [1]
                                                  Prob > chi2     =     0.0000
Log likelihood = -6195.4307                       Pseudo R2       =     0.0049

------------------------------------------------------------------------------
      donate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .8020867   .1051601     7.63   0.000     .5959768    1.008197
       _cons |  -.6647765   .1029084    -6.46   0.000    -.8664731   -.4630798
------------------------------------------------------------------------------
```

Block 2: nuntaught

```
Logistic regression                               Number of obs   =       8999
                                                  LR chi2(2)      =    1314.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -5568.8389                       Pseudo R2       =     0.1056

------------------------------------------------------------------------------
      donate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .8142141    .112313      [2]   0.000     .5940847    1.034344
   nuntaught |   -1.56053   .0457468   -34.11   0.000    -1.650192   -1.470868
       _cons |   .0404119   .1112154     0.36   0.716    -.1775662    .2583901
------------------------------------------------------------------------------
```

Block 3: nuns

```
Iteration 0:   log likelihood = -6226.2249
Iteration 1:   log likelihood = -5563.3657
Iteration 2:   log likelihood =  -5560.861
Iteration 3:   log likelihood = -5560.8602
Iteration 4:   log likelihood = -5560.8602

Logistic regression                                     Number of obs   =       8999
                                                        LR chi2(3)      =    1330.73
                                                        Prob > chi2     =     0.0000
Log likelihood = -5560.8602                             Pseudo R2       =        [3]

------------------------------------------------------------------------------
      donate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .9057642   .1148654     7.89   0.000     .6806322    1.130896
   nuntaught |  -1.569888   .0458944   -34.21   0.000    -1.659839   -1.479936
        nuns |  -.0163351   .0040934    -3.99   0.000    -.0243581   -.0083122
       _cons |  -.0423164   .1132608    -0.37   0.709    -.2643036    .1796707
------------------------------------------------------------------------------


  +-------------------------------------------------------------------+
  | Block |        LL       LR     df  Pr > LR       AIC       BIC |
  |-------+-------------------------------------------------------|
  |     1 | -6195.431    61.59      1   0.0000  12394.86  12409.07 |
  |     2 | -5568.839  1253.18      1   0.0000  11143.68  11164.99 |
  |     3 |  -5560.86    15.96      1   0.0001  11129.72  11158.14 |
  +-------------------------------------------------------------------+
```

Based on the printout above, answer the following.

a. (6 points) Fill in the missing items [1], [2] and [3]. (HINT: The calculations are pretty simple.)

b. (6 pts) Using Model 3 (i.e. Block 3), complete the following table:

male	*nuntaught*	*nuns*	*Log odds*	*Odds*	*P(donate = 1)*
0	0	0			
0	1	0			

c. (9 points) Explain which of the models you think is best, and why. Explain what the model tells us about the effects (or non-effects) of the three independent variables included in the analysis. Be sure to make clear what your preferred model says about the relationship between experience with/approval of nuns and the likelihood of donating.

d. (4 points) The researchers also ran the following:

```
. fre donate

donate
-------------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+----------------------------------------------
Valid   0     |       4273      47.48      47.48      47.48
        1     |       4726      52.52      52.52     100.00
        Total |       8999     100.00     100.00
-------------------------------------------------------------

. estat class

Logistic model for donate

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      3398          1495  |       4893
     -     |      1328          2778  |       4106
-----------+--------------------------+-----------
   Total   |      4726          4273  |       8999

Classified + if predicted Pr(D) >= .5
True D defined as donate != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   71.90%
Specificity                     Pr( -|~D)   65.01%
Positive predictive value       Pr( D| +)   69.45%
Negative predictive value       Pr(~D| -)   67.66%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   34.99%
False - rate for true D         Pr( -| D)   28.10%
False + rate for classified +   Pr(~D| +)   30.55%
False - rate for classified -   Pr( D| -)   32.34%
--------------------------------------------------
Correctly classified                        68.63%
--------------------------------------------------
```

Are you impressed by these results of the classification analysis? Do you think you could have done just as well even without running the logistic regressions?

II-2. (25 points) For <u>each</u> of the following circumstances describe the statistical technique you would use for revealing the relationship between the dependent and independent variables. Write a few sentences explaining and justifying your answer. In some instances more than one technique may be reasonable.

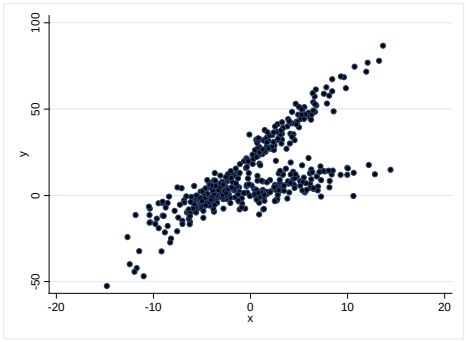
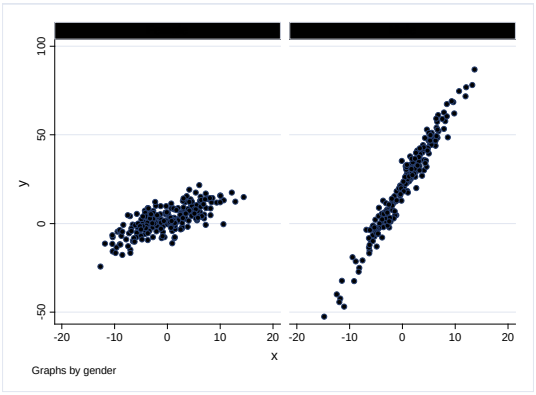
a. A researcher has collected data annually from a panel of respondents for each of the years 2005 thru 2011. Her dependent variable is liberalism measured on a 100 point scale. Unfortunately, some key background variables, such as the political affiliation of the respondent's parents when the respondent was a child, have not been measured.

b. Data have been collected from heterosexual cohabiting couples. Researchers want to know how much the male partner's attitudes about marriage (measured on a 50 point continuous scale) affect the female partner's attitudes about marriage, and vice-versa.

c. A researcher wants to measure the effect of political attitudes on participation in social movements. There are several continuous-level measures of each concept, all of which are believed to suffer from some degree of random measurement error.

d. A professor is teaching two sections of the same class. In one class he is making heavy use of PowerPoint Presentations. In the other class he is lecturing and writing on the board. He wants to see how the two classes compare in terms of student satisfaction, student learning, and classroom attendance.

e. A researcher is interest in how the variables Y, X, and Gender are related. Her scatterplots reveal the following.

`. twoway (scatter y x)`	`. twoway (scatter y x), by(gender )`
	

III. Essay. (30 points) Answer *one* of the following questions.

1. Several assumptions are made when using OLS regression. Discuss TWO of the following in depth. What does the assumption mean? When might the assumption be violated? What effects do violations of the assumption have on OLS estimates? How can violations of the assumption be avoided or dealt with? Be sure to talk about techniques such as 2SLS and logistic regression where appropriate. [NOTE: While the material from the last third of the course is especially relevant here, you should try to tie in earlier material as much as possible too. Also, keep in mind that there are often different ways an assumption can be violated, and the appropriate solutions will therefore often differ too.]

a. The effects of the independent variables are linear and additive
b. Errors are homoskedastic
c. Variables are measured without error
d. All relevant variables are included in the model

2. We've talked about several ways that OLS regression can be modified to deal with violations of its assumptions. Some problems, however, require the use of techniques besides OLS. For three of the following, explain why and when the method would be used instead of OLS. Be sure to make clear what assumptions would be violated if OLS was used instead.

a. 2 stage least squares
b. Logistic regression
c. Robust regression techniques (e.g. rreg, qreg, robust standard errors)
d. Event History Analysis
e. Hierarchical Linear Modeling
f. Ordinal regression

3. Your psychology professor has told you that you should almost always focus on standardized, rather than unstandardized (metric) coefficients. Explain to your professor (as politely as possible) why he is wrong. Among other things, you may want to discuss the relative strengths and weaknesses of standardized vs. unstandardized coefficients with regard to:

a. Variables with arbitrary metrics (e.g. attitudinal scales)
b. Structural equation models
c. Multiple-group comparisons
d. Interpretability of coefficients
e. Effect of random measurement error on coefficients