Sociology 63993
Exam 3 Answer Key
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.

False. Maximum likelihood assumptions about cases being independent are violated with such data, making the use of such things as LR Chi Square tests invalid. Use Wald tests instead.

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.

True. Other possible causes include differing exogenous variances and differing residual variances.

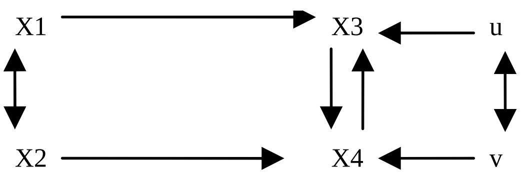
3. The log odds of an event occurring are less than 0. That means that it is impossible for the event to occur.

False. Log odds can vary from negative infinity to positive infinity. If the log odds are negative, this means that the probability of the event occurring are less than 50%.

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

False. While fixed effects models can produce unbiased estimates, they do so by discarding a lot of (possibly biased) information and hence tend to produce larger standard errors.

5. A researcher is interested in the following model:



This model implies that X1 is uncorrelated with X4.

False. X1 is an indirect cause of X4 (X1 affects X3 which in turn affects X4) and X1 is also correlated with one of the causes of X4 (X2). Therefore X1 and X4 are correlated.

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.)

Here are the non-censored parts of the output:

Block 1: male

```
Logistic regression                               Number of obs   =       8999
                                                  LR chi2(1)      =      61.59
                                                  Prob > chi2     =     0.0000
Log likelihood = -6195.4307                       Pseudo R2       =     0.0049
```

Block 2: nuntaught

```
------------------------------------------------------------------------------
      donate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .8142141    .112313     7.25   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

```
Logistic regression                               Number of obs   =       8999
                                                  LR chi2(3)      =    1330.73
                                                  Prob > chi2     =     0.0000
Log likelihood = -5560.8602                       Pseudo R2       =     0.1069
```

To confirm that Stata got it right,

[1] = Likelihood Ratio Chi Square for Model 1 = 61.59. The value is given in the summary table at the end. If you prefer something more challenging, note that LL_0 is included in the output for Block 3 and equals -6226.2249. Ergo, DEV_0 = (-2 * -6226.2249) = 12452.4498, DEV_{M1} = (-2 * -6195.4307) = 12390.8614, So LR Chi Square = $DEV_0 – DEV_{M1}$ = 61.59

[2] = $z_{male} = b_{male}/se_{male}$ = .8142141/.112313 = 7.25

[3] = Pseudo R^2 = Model L^2/ DEV_0 = 1330.73 / 12452.4498 = .1069

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			

Note that the coefficient for nuntaught is -1.569888 and the constant is -.0423164. For the purposes of this problem the other coefficients do not matter because the values of the variables are 0. Ergo,

male	*nuntaught*	*nuns*	*Log odds = a + Xb*	*Odds = exp(LogOdds)*	*P(donate = 1) = Odds/(1 + Odds)*
0	0	0	-.0423164	.9586	.489
0	1	0	-1.6122	.1994	.166

Confirming with Stata,

```
. adjust nuns = 0 male = 0 nuntaught = 0, xb

-------------------------------------------------------------------------------
     Dependent variable: donate     Equation: donate     Command: logit
Covariates set to value: nuns = 0, male = 0, nuntaught = 0
-------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |   -.042316
----------------------
     Key:  xb  =  Linear Prediction
```

```
. adjust nuns = 0 male = 0 nuntaught = 0, exp

---------------------------------------------------------------------------------------
     Dependent variable: donate     Equation: donate     Command: logit
Covariates set to value: nuns = 0, male = 0, nuntaught = 0
---------------------------------------------------------------------------------------

----------------------
      All |    exp(xb)
----------+-----------
          |    .958566
----------------------
     Key:  exp(xb)  =  exp(xb)

. adjust nuns = 0 male = 0 nuntaught = 0, pr

---------------------------------------------------------------------------------------
     Dependent variable: donate     Equation: donate     Command: logit
Covariates set to value: nuns = 0, male = 0, nuntaught = 0
---------------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |    .489422
----------------------
     Key:  pr  =  Probability

. adjust nuns = 0 male = 0 nuntaught = 1, xb

---------------------------------------------------------------------------------------
     Dependent variable: donate     Equation: donate     Command: logit
Covariates set to value: nuns = 0, male = 0, nuntaught = 1
---------------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |    -1.6122
----------------------
     Key:  xb  =  Linear Prediction

. adjust nuns = 0 male = 0 nuntaught = 1, exp

---------------------------------------------------------------------------------------
     Dependent variable: donate     Equation: donate     Command: logit
Covariates set to value: nuns = 0, male = 0, nuntaught = 1
---------------------------------------------------------------------------------------

----------------------
      All |    exp(xb)
----------+-----------
          |    .199448
----------------------
     Key:  exp(xb)  =  exp(xb)
```

```
. adjust nuns = 0 male = 0 nuntaught = 1, pr

---------------------------------------------------------------------------------------
     Dependent variable: donate     Equation: donate     Command: logit
Covariates set to value: nuns = 0, male = 0, nuntaught = 1
---------------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |    .166283
----------------------
     Key:  pr  =  Probability
```

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.

The LR Chi Square contrasts show that Model 3, which includes all three variables, is best. According to the model, men are more likely to donate than are women. But, those who have been taught by nuns, and those who think more highly of nuns, are significantly less likely to donate. As part B showed, in which females with average score on nuns were compared, those taught by nuns were far less likely to donate (16.6%) than those not taught by nuns (48.9%). In the sample, 45% were taught by nuns (not shown), so the Vatican’s actions could be a matter of serious concern for the fundraising efforts.

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?

If you were just guessing, the best approach would be to predict that every case would donate, in which case you would be right 52.52% of the time. The classification table shows that cases were correctly classified 68.63% of the time, which is better than you would do by just guessing.

FYI, if you want to calculate the probability of getting at least 6176 cases out of 8,999 right when p = .5252, the Stata command is

```
. bitesti 8999 6176 0.5252

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
     8999       6176     4726.275       0.52520      0.68630

  Pr(k >= 6176)             = 0.000000  (one-sided test)
  Pr(k <= 6176)             = 1.000000  (one-sided test)
  Pr(k <= 3260 or k >= 6176) = 0.000000  (two-sided test)
```

As Pr(k >= 6176) shows, the probability of being this successful by chance alone is incredibly small. When grading this question, I didn’t care that much whether you thought the analysis was impressive or not, but I did want you to note that the classification table was doing much better than you could expect by chance alone.

II-2. (25 points) For each 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.

Since panel data are available a fixed effects model can be used to control for omitted variables that have time invariant values with time invariant effects. The variable that is specifically mentioned (political affiliation of the respondent's parents when the respondent was a child) seems like it would probably qualify, as might other background variables.

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.

Since there are reciprocal effects and the variables mentioned are continuous, the model is nonrecursive and could be estimated via such means as 2sls or maximum likelihood. (Assuming, of course, that you can specify a model that is identified. This problem sound similar to the Duncan-Haller-Portes model of peer influence that we discussed.)

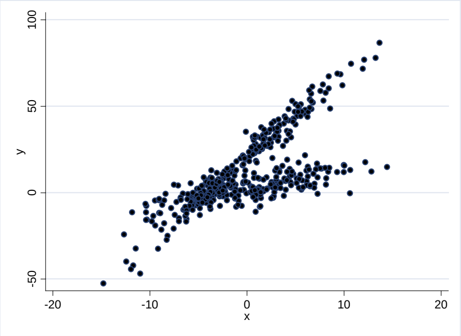
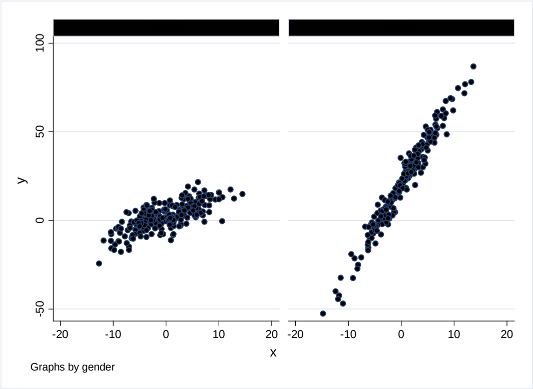
c. A researcher wants to measure the effect of political attitudes on participation in social movements. There are several continuous-level measures of each concept, all of which are believed to suffer from some degree of random measurement error.

The measured variables are continuous and suffer from random measurement error. Using multiple indicators of concepts in a structural equation model (such as can be done by programs like Stata and LISREL) could be a good way of estimating models where the structural effects are not biased by measurement error. An alternative might be to compute scales from the different items (checking reliability via such measures as Cronbach's alpha) and use them instead of the original measures.

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.

The researcher is interested in how multiple outcomes are affected by the same X (PowerPoint vs writing on the board). Manova therefore sounds like a good approach. You could also set this up as a structural equation model.

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 )`
	

It appears that the effect of X on Y differs by gender. You could therefore add the interaction of X * Gender to the model. You could also analyze the groups separately or specify a structural equation model with multiple groups.

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 <u>three</u> 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

See the course notes for ideas on each essay.

Appendix: Stata Code used in the exam

```
* Problem II-1
clear all
webuse nhanes2f
keep in 1/9000
keep female weight heartatk age
drop if missing( age, weight, heartatk, female)

* Create the variables
gen donate = female ==1
gen male = heartatk == 0
gen nuntaught = weight > 72
sum age, meanonly
gen nuns = (age - `=r(mean)') * -1/3
keep  donate male nuntaught nuns

* Run the analysis
nestreg, lr: logit donate male nuntaught nuns
fre donate
estat class

* Confirm the calculations for II-1c
adjust nuns = 0 male = 0 nuntaught = 0, xb
adjust nuns = 0 male = 0 nuntaught = 0, exp
adjust nuns = 0 male = 0 nuntaught = 0, pr
adjust nuns = 0 male = 0 nuntaught = 1, xb
adjust nuns = 0 male = 0 nuntaught = 1, exp
adjust nuns = 0 male = 0 nuntaught = 1, pr

* II-1d Calculate how likely you would do just as well by guessing
bitesti 8999 6176 0.5252
```