Sociology 63993
Exam 3 Answer Key [Draft]
May 10 & 12, 2011

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

1. If a model fails the Brant test, Manova should be used instead.

False. The Brant test is used after ordered logit models. If the model fails the test, you may want to use multinomial logit or generalized ordered logit models instead.

2. The probability of an event occurring is .5. This means that the odds of the event occurring are 1.

True.

3. Unlike OLS regression, a Wald test in logistic regression requires that you estimate both the constrained and unconstrained models.

False. As with OLS, you just estimate the unconstrained model.

4. A model with reciprocal causation is under-identified. One way to solve the problem is to add other variables from the data set that are totally uncorrelated with the variables that are already in the model.

False. The added variables must have direct effects on some variables and indirect effects on others (which means they must be correlated with the variables already in the model.)

5. The dependent variable is coded 1 = Catholic, 2 = Protestant, 3 = Jewish, 4 = Muslim, 5 = Other. Because the DV has more than two categories but is not continuous, an ordered logit model is called for.

False. Categories are not ordered so use mlogit.

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

II-1. (25 points): The New York Times (4/25/2011) recently reported that "Ever since Congress passed the federal gender-equity law known as Title IX, universities have opened their gyms and athletic fields to millions of women who previously did not have chances to play. But as women have surged into a majority on campus in recent years, many institutions have resorted to subterfuge to make it look as if they are offering more spots to women." One of the most questionable practices is that some schools "are counting male practice players as women." For example, Texas A & M, which recently defeated Notre Dame for the national championship in women's basketball, reported that it had 32 players on its team, 14 of whom were men.

Besides raising concerns about whether schools are trying to dodge their legal obligations under Title IX, some coaches are also wondering whether having male practice players on a team actually affects the team's success. Therefore, data have been gathered on every Division I women's sports team in the nation. The variables are

Variable	*Description*
tournament	1 = made the postseason tournament in the team's sport, 0 = did not make tournament
maleplayers	1 = school listed males as part of the women's team, 0 = school did not list males as part of the team
experience	1 = Head Coach has at least 5 years experience, 0 = coach has less than 5 years of experience
salary	Head Coach's salary in thousands of dollars (centered to have a mean of 0)
salary2	Salary Squared

The study obtains the following results:

```
. nestreg, lr: logit  tournament maleplayers experience salary salary2, nolog

Block  1: maleplayers

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(1)      =      13.76
                                                  Prob > chi2     =     0.0002
Log likelihood = -1080.4561                       Pseudo R2       =     0.0063

------------------------------------------------------------------------------
  tournament |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 maleplayers |    .405435    .108947     3.72   0.000     .1919029    .6189671
       _cons |  -1.678001   .0739519   -22.69   0.000    -1.822944   -1.533058
------------------------------------------------------------------------------

Block  2: experience

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(2)      =        [1]
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1041.367                       Pseudo R2       =     0.0423

------------------------------------------------------------------------------
  tournament |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 maleplayers |   .3841133   .1108471     3.47   0.001      .166857    .6013697
  experience |   1.013115   .1197388     8.46   0.000     .7784313    1.247799
       _cons |  -2.292051   .1115846   -20.54   0.000    -2.510753   -2.073349
------------------------------------------------------------------------------

Block  3: salary

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(3)      =     117.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -1028.5212                       Pseudo R2       =        [2]

------------------------------------------------------------------------------
  tournament |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 maleplayers |   .2943249   .1127144      [3]   0.009     .0734088    .5152409
  experience |   1.064714   .1213354     8.77   0.000     .8269012    1.302527
      salary |   .0957424   .0191809     4.99   0.000     .0581486    .1333362
       _cons |   -2.31353   .1131214   -20.45   0.000    -2.535244   -2.091816
------------------------------------------------------------------------------
```

```
Block  4: salary2

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(4)      =     118.31
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1028.182                       Pseudo R2       =     0.0544

------------------------------------------------------------------------------
  tournament |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 maleplayers |   .2954477    .112791     2.62   0.009     .0743814    .5165141
  experience |   1.079981   .1229761     8.78   0.000     .8389525     1.32101
      salary |   .0936368   .0188821     4.96   0.000     .0566286     .130645
     salary2 |   .0032446   .0038837     0.84   0.403    -.0043672    .0108565
       _cons |   -2.35331   .1234685   -19.06   0.000    -2.595304   -2.111316
------------------------------------------------------------------------------


  +-------------------------------------------------------------------+
  | Block |        LL       LR     df  Pr > LR       AIC       BIC |
  |-------+-----------------------------------------------------------|
  |     1 | -1080.456    13.76      1   0.0002  2164.912  2176.388 |
  |     2 | -1041.367    78.18      1   0.0000  2088.734  2105.947 |
  |     3 | -1028.521    25.69      1   0.0000  2065.042  2087.993 |
  |     4 | -1028.182     0.68      1   0.4101  2066.364  2095.052 |
  +-------------------------------------------------------------------+
```

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 is the uncensored printout:

```
Block  2: experience

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(2)      =      91.94
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1041.367                       Pseudo R2       =     0.0423

------------------------------------------------------------------------------
  tournament |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 maleplayers |   .3841133   .1108471     3.47   0.001      .166857    .6013697
  experience |   1.013115   .1197388     8.46   0.000     .7784313    1.247799
       _cons |  -2.292051   .1115846   -20.54   0.000    -2.510753   -2.073349
------------------------------------------------------------------------------

Block  3: salary

Logistic regression                               Number of obs   =       2293
                                                  LR chi2(3)      =     117.63
                                                  Prob > chi2     =     0.0000
Log likelihood = -1028.5212                       Pseudo R2       =     0.0541

------------------------------------------------------------------------------
  tournament |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 maleplayers |   .2943249   .1127144     2.61   0.009     .0734088    .5152409
  experience |   1.064714   .1213354     8.77   0.000     .8269012    1.302527
      salary |   .0957424   .0191809     4.99   0.000     .0581486    .1333362
       _cons |   -2.31353   .1131214   -20.45   0.000    -2.535244   -2.091816
------------------------------------------------------------------------------
```

b. (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 four independent variables included in the analysis. Be sure to make clear what your preferred model says about the relationship between coach's salary and a team's success.

Model 3 fits best. All of the variables are statistically significant. Having male players, an experienced coach, and a coach who has a higher salary makes it more likely the team will make the post-season tournament. The salary2 term is not significant, which implies that it is not eventually counter-productive to keep paying coaches higher and higher salaries.

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

maleplayers	*experience*	*salary*	*Log odds*	*Odds*	*P(tournament = 1)*
1	1	0			
0	1	0			

Using Stata,

```
. quietly logit  tournament maleplayers experience salary
. adjust maleplayers = 1 experience = 1 salary = 0, xb

------------------------------------------------------------------------------
     Dependent variable: tournament     Equation: tournament     Command: logit
Covariates set to value: maleplayers = 1, experience = 1, salary = 0
------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |   -.954491
----------------------
     Key:  xb  =  Linear Prediction

. di exp(-.954491)
.38500806

. adjust maleplayers = 1 experience = 1 salary = 0, pr

------------------------------------------------------------------------------
     Dependent variable: tournament     Equation: tournament     Command: logit
Covariates set to value: maleplayers = 1, experience = 1, salary = 0
------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |    .277983
----------------------
     Key:  pr  =  Probability
```

```
. adjust maleplayers = 0 experience = 1 salary = 0, xb

------------------------------------------------------------------------------
     Dependent variable: tournament     Equation: tournament     Command: logit
Covariates set to value: maleplayers = 0, experience = 1, salary = 0
------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |   -1.24882
----------------------
     Key:  xb  =  Linear Prediction

. di exp(-1.24882)
.28684307

. adjust maleplayers = 0 experience = 1 salary = 0, pr

------------------------------------------------------------------------------
     Dependent variable: tournament     Equation: tournament     Command: logit
Covariates set to value: maleplayers = 0, experience = 1, salary = 0
------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |    .222905
----------------------
     Key:  pr  =  Probability
```

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

```
. fre tournament

tournament
-------------------------------------------------------------
              |      Freq.    Percent      Valid       Cum.
--------------+----------------------------------------------
Valid   0     |       1876      81.81      81.81      81.81
        1     |        417      18.19      18.19     100.00
        Total |       2293     100.00     100.00
-------------------------------------------------------------
```

```
. estat class

Logistic model for tournament

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         0             0  |          0
     -     |       417          1876  |       2293
-----------+--------------------------+-----------
   Total   |       417          1876  |       2293

Classified + if predicted Pr(D) >= .5
True D defined as tournament != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    0.00%
Specificity                     Pr( -|~D)  100.00%
Positive predictive value       Pr( D| +)       .%
Negative predictive value       Pr(~D| -)   81.81%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.00%
False - rate for true D         Pr( -| D)  100.00%
False + rate for classified +   Pr(~D| +)       .%
False - rate for classified -   Pr( D| -)   18.19%
--------------------------------------------------
Correctly classified                        81.81%
--------------------------------------------------
```

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?

Not very impressive. The classification table picked every case to not make the tournament, so it was right 81.81% of the time. You could have been just as successful yourself by picking every case to not make the tournament. Less than 20% of the teams make the tournament and it must be the case that no team has a predicted probability of more than 50%. Any model that doesn't think UConn and Notre Dame have at least a 50% chance of making the women's basketball tournament needs some work!

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 parents' group wants to know what variables affect how much time children spend on Facebook. Five hundred students are asked their age, gender, and whether or not they have a phone that can access Facebook. They are also asked to indicate whether they never use Facebook, spend less than an hour a day on Facebook, or spend more than an hour a day on Facebook.

The dependent variable is ordinal, so an ordered logit model appears most appropriate, or perhaps an mlogit model if the ordered logit assumptions are violated.

b. Two weeks ago, three hundred people were asked to indicate, on a 100 point scale, how much they approved or disapproved of the job Barack Obama was doing as President. Two days after the death of Osama bin Laden, those same people were again asked how much they approved or disapproved of the job Barack Obama was doing as President.

A matched-pairs t-test is appropriate. Subjects are tested both before and after Osama's death.

c. A campaign manager observes that, between the ages of 20 to 40, people who are older tend to be slightly more supportive of her candidate. However, after age 40, the positive effect of age becomes much greater.

You should probably have a model with a spline function, where the slope changes (and becomes steeper) after 40.

d. A group of educators has developed a remedial reading program. The group feels that, the longer a student sticks with the program the better his or her reading scores will be. They are concerned, however, by the large numbers of students who drop out of the program somewhere along the way. Therefore, for one thousand students, it has collected data on how many weeks the student stayed with the program before dropping out (if ever). It has also collected data on each student's age, race, socio-economic status, and whether or not the student lives in a female-headed household.

Event history analysis sounds good. People can drop out at different times (or not drop out at all) so EHA will let you examine what factors speed up or slow down the pace of dropping out.

e. A researcher is interested in the effects of education, race, IQ, and gender on Income. Due to problems in data collection, IQ was not determined for 15% of the sample. Even though these cases are believed to be missing at random, the researcher is not happy about the prospect of losing 15% of the cases in her analysis.

Multiple imputation would be good. You can impute values for the missing cases, hence keeping all of the cases in the analysis.

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. Ordered Logit models
d. Robust regression techniques (e.g. rreg, qreg, robust standard errors)
e. Event History Analysis
f. Hierarchical Linear Modeling

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

```
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
* Create the variables
gen tournament = warm==4
gen maleplayers = yr89
gen experience = male ==0
center ed
clonevar salary = c_ed
gen salary2 = salary ^ 2
* Run the analysis
nestreg, lr: logit  tournament maleplayers experience salary salary2, nolog
fre tournament
estat class
* Confirm the log odds, odds, & probabilities
quietly logit  tournament maleplayers experience salary
adjust maleplayers = 1 experience = 1 salary = 0, xb
di exp(-.954491)
adjust maleplayers = 1 experience = 1 salary = 0, pr
adjust maleplayers = 0 experience = 1 salary = 0, xb
di exp(-1.24882)
adjust maleplayers = 0 experience = 1 salary = 0, pr
```