Sociology 63993
Exam 3 Answer Key
May 4 & 7, 2010

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

1. The use of standardized coefficients is one way to effectively deal with problems of random measurement error in variables.

False. If anything standardization will make problems worse. When random variability is added to the dependent variable, its variance increases and the standardized effects go down (because their computation involves the variance of the DV). Estimates of the underlying structural relationships (i.e. the non-standardized coefficients) are not biased by random measurement error in the DV though.

2. Robust regression techniques should be used to estimate nonrecursive models.

False. 2sls is a method for estimating nonrecursive models.

3. The log odds of an event occurring are 0. This means that the event cannot happen.

False. If the log odds are 0, the odds are 1, and the probability of the event occurring is 50 percent.

4. Y is regressed on X in two different populations. In both populations, the estimated Beta coefficient equals 3. This means that the R^2 value will also be the same in the two populations.

False. The value of R^2 is also affected by the variance of the exogenous variable and the residual variance.

5. After running an ordered logit model, a researcher obtains the following:

```
. brant

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |     49.18    0.000    12
-------------+--------------------------
        yr89 |     13.01    0.001     2
        male |     22.24    0.000     2
       white |      1.27    0.531     2
         age |      7.38    0.025     2
          ed |      4.31    0.116     2
        prst |      4.33    0.115     2
----------------------------------------
```

These results suggest that the assumptions of the model have been violated in this analysis.

True. If the assumptions are not violated you will get small chi-square values. Violations are especially severe for male and yr89.

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

II-1. (25 points): After several setbacks in recent elections, supporters of gay marriage are cautiously optimistic about their prospects for the future. A small but growing number of states have legalized gay marriage. In January 2010, Ted Olson (George Bush's attorney in Bush versus Gore 2000) surprised many by writing an article for Newsweek entitled *The conservative case for gay marriage*. A February 2010 study by the Pew Research Center found that young adults were much less opposed to gay marriage than were older generations.

Still, gay rights supporters know there is a long struggle ahead. They want to better understand the factors that influence attitudes toward gay marriage. On the one hand, they believe that having gay friends, or living in a state that has legalized gay marriage, will lead to more favorable attitudes. At the same time, they believe that certain demographic characteristics, such as race and age, will also have an effect. They have therefore collected data from more than 10,000 American adults on the following:

Variable	*Description*
gaymarr	1 = supports gay marriage, 0 = opposes it
gayfriends	1 = has 1 or more gay friends, 0 = no gay friends
legalstate	1 = lives in a state where gay marriage is legal, 0 = gay marriage is not legal in the state
black	1 = black, 0 = not black
agecentered	Age of respondent (centered to have a mean of 0)

They obtain the following results (some extraneous output is deleted):

```
. nestreg, lr: logit gaymarr gayfriends legalstate (agecentered black), nolog

Block  1: gayfriends

Logistic regression                               Number of obs   =      10335
                                                  LR chi2(1)      =    1493.81
                                                  Prob > chi2     =     0.0000
Log likelihood = -6403.8352                       Pseudo R2       =     0.1045

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gayfriends |   1.585525   .0426173    37.20   0.000     1.501996    1.669053
       _cons |  -.8340089   .0290552   -28.70   0.000     -.890956   -.7770617
------------------------------------------------------------------------------

Block  2: legalstate

Logistic regression                               Number of obs   =      10335
                                                  LR chi2(2)      =    1555.20
                                                  Prob > chi2     =     0.0000
Log likelihood = -6373.1378                       Pseudo R2       =     0.1087

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gayfriends |   1.584484   .0427574    37.06   0.000     1.500681    1.668287
  legalstate |   .8155892   .1065616    [ 1 ]   0.000      .6067323    1.024446
       _cons |  -.8697655   .0295722   -29.41   0.000    -.9277259   -.8118052
------------------------------------------------------------------------------
```

Block 3: agecentered black

```
Logistic regression                                     Number of obs   =      10335
                                                        LR chi2(4)      =    1584.47
                                                        Prob > chi2     =     0.0000
Log likelihood = -6358.5032                             Pseudo R2       =      [ 2 ]

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gayfriends |    1.60257   .0430369    37.24   0.000     1.518219    1.686921
  legalstate |   .9172599   .1091589     8.40   0.000     .7033122    1.131207
 agecentered |   -.005885   .0012709    -4.63   0.000    -.0083758   -.0033941
       black |  -.2050577   .0697825    -2.94   0.003    -.3418289   -.0682864
       _cons |  -.8614515   .0303248   -28.41   0.000    -.9208871    -.802016
------------------------------------------------------------------------------

  +-------------------------------------------------------------------+
  | Block |        LL       LR     df  Pr > LR       AIC       BIC |
  |-------+-------------------------------------------------------|
  |     1 | -6403.835  1493.81      1   0.0000  12811.67  12826.16 |
  |     2 | -6373.138    61.39      1   0.0000  12752.28  12774.01 |
  |     3 | -6358.503    [ 3 ]      2   0.0000  12727.01  12763.22 |
  +-------------------------------------------------------------------+
```

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. Previously omitted values are in ***bold italics***.

```
Block  2: legalstate

Logistic regression                                     Number of obs   =      10335
                                                        LR chi2(2)      =    1555.20
                                                        Prob > chi2     =     0.0000
Log likelihood = -6373.1378                             Pseudo R2       =     0.1087

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gayfriends |   1.584484   .0427574    37.06   0.000     1.500681    1.668287
  legalstate |   .8155892   .1065616     7.65   0.000     .6067323    1.024446
       _cons |  -.8697655   .0295722   -29.41   0.000    -.9277259   -.8118052
------------------------------------------------------------------------------

Block  3: agecentered black

Logistic regression                                     Number of obs   =      10335
                                                        LR chi2(4)      =    1584.47
                                                        Prob > chi2     =     0.0000
Log likelihood = -6358.5032                             Pseudo R2       =     0.1108

------------------------------------------------------------------------------
     gaymarr |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  gayfriends |    1.60257   .0430369    37.24   0.000     1.518219    1.686921
  legalstate |   .9172599   .1091589     8.40   0.000     .7033122    1.131207
 agecentered |   -.005885   .0012709    -4.63   0.000    -.0083758   -.0033941
       black |  -.2050577   .0697825    -2.94   0.003    -.3418289   -.0682864
       _cons |  -.8614515   .0303248   -28.41   0.000    -.9208871    -.802016
------------------------------------------------------------------------------
```

Block	LL	LR	df	Pr > LR	AIC	BIC
1	-6403.835	1493.81	1	0.0000	12811.67	12826.16
2	-6373.138	61.39	1	0.0000	12752.28	12774.01
3	-6358.503	***29.27***	2	0.0000	12727.01	12763.22

To confirm that Stata got it right:

[1] = $z_{legalstate} = b_{legalstate}/se_{legalstate} = .8155892/.1065616 = 7.65$.

[2] = Pseudo $R^2 = G_M/ (G_M + -2*LL_M) = 1584.47/ (1584.47 + 2*6358.5032) = .11108$

[3] = Incremental LR for Model 3 = $G_3 – G_2 = 1584.47 – 1555.20 = 29.27$

b. (7 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.

Both the chi-square contrasts and the z values show that Model 3 provides the best fit to the data, with all of the variables having significant effects. According to this model, those with gay friends and those who live in a state where gay marriage is legal tend to be more likely to support gay marriage. This is consistent with what the researchers hypothesized. Conversely, blacks and older people tend to be less supportive.

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

Gayfriends	*Legalstate*	*Black*	*Agecentered*	*Log odds*	*Odds*	*P(supporting gay marriage))*
Has gay friends	Gay marriage is legal in R's state	0	0			
Does not have gay friends	Gay marriage is not legal in R's state	0	0			

Note that the coefficient for gayfriends is 1.60257, the coefficient for legalstate = .9172599, and the constant is -.8614515. For the purposes of this problem, the other coefficients do not matter because the values of the variables are 0. Ergo,

Gayfriends	*Legalstate*	*Black*	*Agecentered*	*Log odds = a + XB*	*Odds = exp(LogOdds)*	*P(Supporting Gay Marriage) = Odds/(1 + Odds)*
Has gay friends	Gay marriage is legal in R's state	0	0	1.6583784	5.2508	.84
Does not have gay friends	Gay marriage is not legal in R's state	0	0	-.8614515	.4225	.297

We can confirm with Stata:

```
. adjust agecentered = 0 black = 0 legalstate = 1 gayfriends = 1, xb

------------------------------------------------------------------------------
     Dependent variable: gaymarr     Equation: gaymarr     Command: logit
Covariates set to value: agecentered = 0, black = 0, legalstate = 1, gayfriends = 1
------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |    1.65838
----------------------
     Key:  xb  =  Linear Prediction

. adjust agecentered = 0 black = 0 legalstate = 1 gayfriends = 1, exp

------------------------------------------------------------------------------
     Dependent variable: gaymarr     Equation: gaymarr     Command: logit
Covariates set to value: agecentered = 0, black = 0, legalstate = 1, gayfriends = 1
------------------------------------------------------------------------------

----------------------
      All |    exp(xb)
----------+-----------
          |    5.25079
----------------------
     Key:  exp(xb)  =  exp(xb)

. adjust agecentered = 0 black = 0 legalstate = 1 gayfriends = 1, pr

------------------------------------------------------------------------------
     Dependent variable: gaymarr     Equation: gaymarr     Command: logit
Covariates set to value: agecentered = 0, black = 0, legalstate = 1, gayfriends = 1
------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |     .84002
----------------------
     Key:  pr  =  Probability
```

```
. adjust agecentered = 0 black = 0 legalstate = 0 gayfriends = 0, xb

------------------------------------------------------------------------------
     Dependent variable: gaymarr     Equation: gaymarr     Command: logit
Covariates set to value: agecentered = 0, black = 0, legalstate = 0, gayfriends = 0
------------------------------------------------------------------------------

----------------------
      All |         xb
----------+-----------
          |   -.861452
----------------------
     Key:  xb  =  Linear Prediction

. adjust agecentered = 0 black = 0 legalstate = 0 gayfriends = 0, exp

------------------------------------------------------------------------------
     Dependent variable: gaymarr     Equation: gaymarr     Command: logit
Covariates set to value: agecentered = 0, black = 0, legalstate = 0, gayfriends = 0
------------------------------------------------------------------------------

----------------------
      All |    exp(xb)
----------+-----------
          |    .422548
----------------------
     Key:  exp(xb)  =  exp(xb)

. adjust agecentered = 0 black = 0 legalstate = 0 gayfriends = 0, pr

------------------------------------------------------------------------------
     Dependent variable: gaymarr     Equation: gaymarr     Command: logit
Covariates set to value: agecentered = 0, black = 0, legalstate = 0, gayfriends = 0
------------------------------------------------------------------------------

----------------------
      All |         pr
----------+-----------
          |    .297036
----------------------
     Key:  pr  =  Probability
```

d. (6 points) The researchers also ran the following:

```
. tab1 gaymarr

-> tabulation of gaymarr

    gaymarr |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,426       52.50       52.50
          1 |      4,909       47.50      100.00
------------+-----------------------------------
      Total |     10,335      100.00
```

```
. quietly logit gaymarr gayfriends legalstate (agecentered black)

. estat class

Logistic model for gaymarr

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      3221          1526  |       4747
     -     |      1688          3900  |       5588
-----------+--------------------------+-----------
   Total   |      4909          5426  |      10335

Classified + if predicted Pr(D) >= .5
True D defined as gaymarr != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   65.61%
Specificity                     Pr( -|~D)   71.88%
Positive predictive value       Pr( D| +)   67.85%
Negative predictive value       Pr(~D| -)   69.79%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   28.12%
False - rate for true D         Pr( -| D)   34.39%
False + rate for classified +   Pr(~D| +)   32.15%
False - rate for classified -   Pr( D| -)   30.21%
--------------------------------------------------
Correctly classified                        68.90%
--------------------------------------------------
```

Are you impressed by these results of the classification analysis? Do you think you could have done just as well by randomly guessing whether someone supported gay marriage or not?

If you were just guessing, you could guarantee that you would be right 52.5% of the time by always predicting that the person would oppose gay marriage. Anything better than that would require luck. The classification table gets 68.9% of the cases (i.e. 7,121 cases out of 10,335 total) classified correctly, so it is doing quite a bit better than you would expect by chance alone. The classification table tends to have little value when there are extreme splits in either direction, but is somewhat more useful when the split is more in the 50-50 range.

FYI, if you want to calculate the probability of getting at least 7,121 cases out of 10,335 right when p = .525, the Stata command is

```
. bitesti 10335 7121 0.525

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
    10335       7121     5425.875       0.52500      0.68902

  Pr(k >= 7121)              = 0.000000  (one-sided test)
  Pr(k <= 7121)              = 1.000000  (one-sided test)
  Pr(k <= 3711 or k >= 7121) = 0.000000  (two-sided test)
```

As Pr(k >= 7121) 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. It is November 2012. Polls show that the Tea Party Presidential ticket of Sarah Palin and Glenn Beck is within striking distance of pulling off the greatest upset in American electoral history. Palin's campaign knows that it must persuade the nation's few remaining undecided voters if it hopes to win. The campaign has therefore prepared two ads. In the first ad, Palin vows to eliminate wasteful spending, such as all forms of federal support for graduate student education. In the second ad, Palin is shown hunting in Alaska and her longtime support for gun owner rights is stressed. Each ad will be shown to a different group of 500 undecided voters. Using a 100 point scale, voters will be asked how much they liked the ad they saw. Whichever ad is found to be most effective will air on election eve.

A simple T-test is called for. The dependent variable is effectiveness of ad, and the independent variable is the type of ad seen. You could also do this via ANOVA or by regressing effectiveness on type of ad.

b. As luck would have it, British Petroleum conducted a survey on April 15th in which respondents were asked about their support for offshore drilling, their attitudes toward oil company profits, and whether they thought oil companies needed to be regulated more. All three attitudes were measured on continuous scales. Naturally, the company is worried about how the subsequent Gulf oil spill is going to affect public opinion. It will therefore conduct another survey on May 15 that asks these same questions.

There are three dependent variables, so MANOVA is called for. The independent variable is code 0 = before spill, 1 = after spill. You could also set this up in a program like LISREL.

c. Stata Corporation is thrilled by increased sales of its product and wants to determine how it can get academics to start using its software sooner. One thousand academics who do not use Stata (but who have access to it on their campus network) are surveyed monthly over a three year period. Independent variables include the amount of Stata advertising material the academic received, the number of personal phone calls made by Stata marketing representatives, and the number of times Stata was cited during each monthly period in the three journals the academic reads most. The date on which the academic first started using Stata (if ever) is also recorded.

Event history analysis is called for. Stata is interested in what affects the rate at which Stata is adopted; it isn't simply interested in whether or not Stata gets adopted, it wants to know what makes some people adopt Stata sooner than other people do. The independent variables are advertising material, citations, etc. Note also that the dependent variable is right-censored. Some people may adopt Stata after the three year study is over.

d. A researcher adamantly believes that attitudes toward feminism are strong determinants of how much someone likes FaceBook. She is therefore amazed to discover that not one of the 300 previous studies that have tested her hypothesis have found one shred of evidence to support it. She thinks that this must surely be due to problems of poor measurement. She has therefore written three questions that all tap feminist attitudes and another three questions that all measure liking for FaceBook. All the items use continuous scales. All six items will be asked as part of a national survey of 1200 respondents.

This would be a good problem for LISREL or a related program. There are two underlying latent variables (feminism and liking for FaceBook) each of which has three observed but flawed indicators. You could also consider creating two sets of scales from the items and then use regular OLS.

e. A professor has just spent a small fortune fixing plumbing and electrical problems in his house. He is therefore amazed to learn that several graduate students have bought or are considering buying their own homes. To find out why in the world they would want to do this, he has prepared a short questionnaire that will be administered to 100 randomly

selected students. Students will be asked if they agree or disagree with the following statement: "It is a good idea for graduate students like me to own their own homes." Possible responses are Strongly Agree, Agree, Disagree and Strongly Disagree. They will also be asked questions about their income, whether their parents owned or rented their home, how many years they expect to be in graduate school, whether they have a spouse or partner who is employed, and whether or not they are aware of government programs that could help them to finance a house.

The dependent variable ("It is a good idea…") is ordinal, so some sort of ordinal regression model is probably called for, e.g. ordered logit. If the assumptions of the model are violated some other type of ordinal method may have to be used, or the researcher may have to use the less parsimonious mlogit method.

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. Path analysis first became popular in Sociology during the 1960s, and has evolved considerably since then.

a. In the early days of path analysis, standardized coefficients were widely used. Give two or three reasons why, in Sociology at least, that practice fell out of favor.

b. In the 1970s, the development of the LISREL program gave new life to path analysis. Discuss some of the key strengths of the LISREL method. Explain how LISREL made it possible to estimate important new sorts of models and how it provided an alternative means for estimating models that could also be approached via other methods.

See the course notes for ideas on each essay.

Appendix: Stata Code used in the exam

```
* Problem I-5
use "http://www.indiana.edu/~jslsoc/stata/spex_data/ordwarm2.dta", clear
quietly ologit  warm yr89 male white age ed prst
brant

* Problem II-1
clear all
webuse nhanes2f
keep female weight black heartatk age
drop if missing( age, weight, heartatk, female, black)
* Create the variables
gen gaymarr = female ==0
clonevar legalstate = heartatk
gen gayfriends = weight > 72
sum age, meanonly
gen agecentered = age - `=r(mean)'
keep  gaymarr legalstate gayfriends black age agecentered
* Run the analysis
nestreg, lr: logit gaymarr gayfriends legalstate (agecentered black), nolog
tab1 gaymarr
quietly logit gaymarr gayfriends legalstate (agecentered black)
estat class
* Confirm the calculations for II-1c
adjust agecentered = 0 black = 0 legalstate = 1 gayfriends = 1, xb
adjust agecentered = 0 black = 0 legalstate = 1 gayfriends = 1, exp
adjust agecentered = 0 black = 0 legalstate = 1 gayfriends = 1, pr
adjust agecentered = 0 black = 0 legalstate = 0 gayfriends = 0, xb
adjust agecentered = 0 black = 0 legalstate = 0 gayfriends = 0, exp
adjust agecentered = 0 black = 0 legalstate = 0 gayfriends = 0, pr
* II-1d Calculate how likely you would do just as well by guessing
bitesti 10335 7121 0.525
```