Sociology 63993, Exam 3 Answer Key
May 1 and May 6, 2015

Richard Williams, University of Notre Dame, http://www3.nd.edu/~rwilliam/

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

1. In a logistic regression the Pseudo R^2 is .5. This means that half the respondents experience the event.

False. The pseudo R^2 gives you an idea of how strong the association is between the dependent and independent variables, but tells you nothing about the split between 1s and 0s.

2. One reason some people do not like random effects models is that they tend to have much larger standard errors than do fixed effects models.

False. The opposite is true. Random effects models tend to have lower standard errors – but their coefficients are more likely to be biased.

3. Y is regressed on X in two different populations. In both populations, the variance of the disturbance term equals 3. This means that the R^2 value will also be the same in the two populations.

False. The residual variance is just one of the three factors that influence R^2. The structural coefficient and the exogenous variance are also important.

4. The dependent variable Y suffers from random measurement error. Therefore, when doing cross-population comparisons, it is best to focus on the standardized coefficients.

False. Random measurement error in Y does not bias the metric coefficients but it does bias the standardized coefficients.

5. A physician has developed a new exercise program. She believes that those who participate in the program will be happier, more physically fit, and will work better on the job than those who do not participate. Happiness, physical fitness, and job productivity are all measured on interval-level scales. Participation in the program is coded 0 or 1. Her best strategy is to simply run three different OLS regressions.

False. Since there are multiple dependent variables MANOVA would be a better choice.

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

II-1. (25 points): It is September 2016. After his stunning and decisive upset victory over Jeb Bush in the Indiana primary, Republican Presidential candidate Ted Cruz now faces the daunting task of taking on heavily favored Hillary Clinton. Cruz, however, remains optimistic. First, he believes it is actually a very close race at the moment. Further, if he can identify which of his issues resonates most with the American people, he is confident he can win and provide the nation with the change in leadership it so desperately needs. His pollsters have therefore gathered the following information from over 4,000 likely voters:

Variable	*Description*
cruz	1 = supports Cruz, 0 = does not support Cruz
male	1 = male, 0 = female
tradmar	Supports traditional marriage and opposes gay marriage. 1 = opposes gay marriage, 0 = supports gay marriage
fiscalconserv	Fiscal conservatism scale. The higher the score, the more fiscally conservative the respondent is. The scale has been centered to have a mean of zero.

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

```
. fre cruz

cruz
----------------------------------------------------------------------
                        |      Freq.    Percent      Valid       Cum.
------------------------+---------------------------------------------
Valid   0 Opposes Cruz  |       2649      63.60      63.60      63.60
        1 Supports Cruz |       1516      36.40      36.40     100.00
        Total           |       4165     100.00     100.00
----------------------------------------------------------------------

. nestreg, lr: logit cruz male tradmar fiscalcons, nolog

Block  1: male

Logistic regression                             Number of obs     =      4,165
                                                LR chi2(1)        =        [1]
                                                Prob > chi2       =     0.0000
Log likelihood =  -2421.107                     Pseudo R2         =     0.1134

------------------------------------------------------------------------------
        cruz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.699879   .0724706    23.46   0.000     1.557839    1.841919
       _cons |  -1.531327   .0579638   -26.42   0.000    -1.644934    -1.41772
------------------------------------------------------------------------------

Block  2: tradmar

Logistic regression                             Number of obs     =      4,165
                                                LR chi2(2)        =     655.48
                                                Prob > chi2       =     0.0000
Log likelihood = -2403.1552                     Pseudo R2         =     0.1200

------------------------------------------------------------------------------
        cruz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.680128   .0727543      [2]   0.000     1.537532    1.822723
     tradmar |   .7221417   .1253889     5.76   0.000      .476384    .9678994
       _cons |  -2.171609   .1282316   -16.94   0.000    -2.422939    -1.92028
------------------------------------------------------------------------------

Block  3: fiscalcons

Logistic regression                             Number of obs     =      4,165
                                                LR chi2(3)        =     718.70
                                                Prob > chi2       =     0.0000
Log likelihood = -2371.5439                     Pseudo R2         =     0.1316

------------------------------------------------------------------------------
        cruz |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   1.931706   .0815088    23.70   0.000     1.771951     2.09146
     tradmar |   .3492499   .1340515     2.61   0.009     .0865137    .6119861
  fiscalcons |   .1392001   .0176962     7.87   0.000     .1045162    .1738841
       _cons |   -1.98021   .1298305   -15.25   0.000    -2.234673   -1.725747
------------------------------------------------------------------------------
```

```
+-------------------------------------------------------------+
| Block |        LL       LR     df  Pr > LR       AIC       BIC |
|-------+-----------------------------------------------------|
|     1 | -2421.107   619.58      1   0.0000  4846.214  4858.883 |
|     2 | -2403.155    35.90      1   0.0000   4812.31  4831.314 |
|     3 | -2371.544     [3]       1   0.0000  4751.088  4776.426 |
+-------------------------------------------------------------+
```

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.)

[1] = Model chi-square for model 1= 619.58. The number is already reported in the summary table for model 1.
[2] = $z_{male} = b_{male}/ se_{male}$ = 1.680128/ .0727543 = 23.09
[3] = Incremental chi-square for model 3 = LR chi-square for model 3 – LR chi-square for model 2 = 718.70 – 655.48 = 63.22

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

male	*tradmar*	*fiscalcons*	*Log odds*	*Odds*	*P(cruz = 1)*
0	0	0			
0	1	0			

Note that the coefficient for tradmar is .3492499 and the constant is -1.98021. For the purposes of this problem the other coefficients do not matter because the values of the variables are 0. Ergo,

male	*tradmar*	*fiscalcons*	*Log odds = a + Xb*	*Odds = exp(LogOdds)*	*P(cruz = 1) = Odds/(1 + Odds)*
0	0	0	-1.98021	.1380403	.1212965
0	1	0	-1.63096	.1957416	.1636989

The margins command in Stata can do this easily, especially if we redo the logit command using factor variables:

```
. quietly logit cruz i.male i.tradmar fiscalcons
. * Log Odds
. margins tradmar, at(male = 0 fiscalcons = 0) predict(xb)

Adjusted predictions                            Number of obs     =      4,165
Model VCE    : OIM

Expression   : Linear prediction (log odds), predict(xb)
at           : male            =           0
               fiscalcons      =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tradmar |
          0  |   -1.98021   .1298305   -15.25   0.000    -2.234673   -1.725747
          1  |   -1.63096   .0652333   -25.00   0.000    -1.758815   -1.503105
------------------------------------------------------------------------------

. * Odds
. margins tradmar, at(male = 0 fiscalcons = 0) expression(exp(predict(xb)))

Adjusted predictions                            Number of obs     =      4,165
Model VCE    : OIM

Expression   : exp(predict(xb))
at           : male            =           0
               fiscalcons      =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tradmar |
          0  |   .1380403   .0179218     7.70   0.000     .1029141    .1731664
          1  |   .1957416   .0127689    15.33   0.000      .170715    .2207681
------------------------------------------------------------------------------

. * Probabilities
. margins tradmar, at(male = 0 fiscalcons = 0)

Adjusted predictions                            Number of obs     =      4,165
Model VCE    : OIM

Expression   : Pr(cruz), predict()
at           : male            =           0
               fiscalcons      =           0

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tradmar |
          0  |   .1212965   .0138378     8.77   0.000     .0941749    .1484181
          1  |   .1636989   .0089305    18.33   0.000     .1461953    .1812024
------------------------------------------------------------------------------
```

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. Also tell Cruz whether he is ahead or behind at this point.

All variables are significant in model 3. Men, those who oppose gay marriage, and people who are more fiscally conservative are more like to support Cruz than are others. Alas, he still trails badly, as the frequency shows that only 36.4% of the population currently supports him.

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

```
. estat class

Logistic model for cruz

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |      1021           552  |       1573
     -     |       495          2097  |       2592
-----------+--------------------------+-----------
   Total   |      1516          2649  |       4165

Classified + if predicted Pr(D) >= .5
True D defined as cruz != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   67.35%
Specificity                     Pr( -|~D)   79.16%
Positive predictive value       Pr( D| +)   64.91%
Negative predictive value       Pr(~D| -)   80.90%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   20.84%
False - rate for true D         Pr( -| D)   32.65%
False + rate for classified +   Pr(~D| +)   35.09%
False - rate for classified -   Pr( D| -)   19.10%
--------------------------------------------------
Correctly classified                        74.86%
--------------------------------------------------
```

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? Put another way, are more cases correctly classified by the logistic regression than you likely would have correctly classified yourself?

Based on the frequencies, the best strategy would be to always pick against Cruz, as you would be right 63.6% of the time, or about 2,649 times. But, the table correctly classified 74.86% of the respondents, 3,1118 cases, which is quite a bit better. The bitesti command confirms that it would be almost impossible to do this well by luck alone:

```
. bitesti 4165 3118 0.636, detail

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
     4165       3118      2648.94       0.63600      0.74862

  Pr(k >= 3118)              = 0.000000  (one-sided test)
  Pr(k <= 3118)              = 1.000000  (one-sided test)
  Pr(k <= 2158 or k >= 3118) = 0.000000  (two-sided test)

  Pr(k == 3118)              = 0.000000  (observed)
  Pr(k == 2159)              = 0.000000
  Pr(k == 2158)              = 0.000000  (opposite extreme)
```

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. Some problems may require the use of advanced techniques while in other instances the required technique may be simple and basic.

a. A researcher has collected data from the same set of respondents annually for each of the last five years. She now realizes that the age of the respondent's mother when the respondent was born needs to be incorporated into her models or at least controlled for in some way. Unfortunately this variable was not measured in her surveys.

A fixed effects regression model could be good as such models control for the effects of time invariant variables that have time invariant effects. (Of course, if she is going to be collecting more data, she could just ask the question during the next wave of the survey.)

b. In both 2004 and 2014 Notre Dame students were asked how supportive they were of gay and lesbian rights. The scale ranged from a low of 1 (very unsupportive) to a high of 100 (extremely supportive). The newly formed campus group OUTatND has gained access to the data and hypothesizes that support is greater now than it was 10 years ago.

A simple T test of independent samples would be fine. If you want to be fancier you could run a bivariate regression with a dummy variable for year.

c. Hillary Clinton's campaign has gathered data on 5 items that they think measure political liberalism and another 5 items that they think measure support for Clinton. They believe that the more liberal someone is, the stronger their support will be for Clinton. However they are concerned about how accurate their estimates will be since all 10 items are believed to suffer from random measurement error.

A structural equation model could be good. You could specify two latent variables, each of which had 5 indicators. If the model was good this could control for the effects of random measurement error and give an unbiased estimate of the relationship between the two underlying variables.

d. The National Rifle Association has collected data from both wives and their husbands. Each spouse has been asked to rank their support for gun control laws on a scale that runs from 0 to 50. The NRA believes that husbands and wives influence each other, i.e. the husband's attitude on gun control affects his wife's attitude and the wife's attitude affects her husband's attitude.

Assuming you could get it identified, a non-recursive model seems called for. You could use 2sls or structural equation modeling.

e. Educational researchers are trying to determine the optimal amount of homework to give to 8th grade students. They believe that too little homework results in too little learning, and so at least some homework should be given. But, at the same time, they worry that, after a certain point, if students are given too much homework then learning will start to decline. They have data from over 10,000 8th graders nationwide that includes information on the amount of homework given and the amount students learned.

This sounds like a curvilinear relationship, so including homework^2 in the model would probably be a good idea. Alternatively you could consider a spline model if you had a good idea of where the turning point is.

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. Fixed effects regression models
f. Structural Equation Modeling using multiple indicators of variables

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

Appendix: Stata Code used in the exam

```
version 13.1
* II-1: Ted Cruz problem
use "http://statisticalhorizons.com/wp-content/uploads/wages.dta", clear
* Set up data
gen cruz = union
label define cruz 0 "Opposes Cruz" 1 "Supports Cruz"
label values cruz cruz
gen male = occ
gen tradmar = fem==0
gen fiscalcons = lwage * 5
center fiscalcons, inplace
keep cruz male tradmar fiscalcons
* Run analyses
fre cruz
nestreg, lr: logit cruz male tradmar fiscalcons, nolog
estat clas
*Supplemental analyses
quietly logit cruz i.male i.tradmar fiscalcons
* Log Odds
margins tradmar, at(male = 0 fiscalcons = 0) predict(xb)
* Odds
margins tradmar, at(male = 0 fiscalcons = 0) expression(exp(predict(xb)))
* Probabilities
margins tradmar, at(male = 0 fiscalcons = 0)
* Lielihook of classifying this many correctly via luck
bitesti 4165 3118 0.636, detail
```