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

1. \( R^2 = .5 \) in both populations A and B. This means that the structural effect of X is the same in both populations.

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

3. A researcher obtains the following:

```
. logit warmlt2 age, or nolog
Logistic regression                       Number of obs = 2293
LR chi2(1) = 36.17                       Prob > chi2 = 0.0000
Log likelihood = -865.82744             Pseudo R2 = 0.0205

+----------------------------------------+----------+----------------+-----------------+-----------------+------------------+
<p>| warmlt2 | Odds Ratio | Std. Err. | z     | P&gt;|z| | [95% Conf. Interval] |
|---------|------------|-----------|-------|-----|------------------------|</p>
<table>
<thead>
<tr>
<th>age</th>
<th>1.022106</th>
<th>.0037148</th>
<th>6.02</th>
<th>0.000</th>
<th>1.014851 1.029412</th>
</tr>
</thead>
</table>
```

This means that, with each additional year of age, the log odds increase by slightly over 1.

4. In logistic regression, as in OLS regression, Stata commands like `collin` can be used to test for multicollinearity.

5. A researcher wants to test the hypothesis

\[ H_0: \mu_{11} = \mu_{21} \]
\[ \mu_{12} = \mu_{22} \]
\[ \mu_{13} = \mu_{23} \]

A nonrecursive model is called for.

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

II-1. (25 points) A medical researcher and a sociologist have teamed up to do work on the relationship between health, race, residence, and concentrations of lead in the body. The sociologist stresses that race and residence are very important. The medical researcher agrees and adds that the amount of lead in the body is very important. They therefore collect information on the following variables:
<table>
<thead>
<tr>
<th><strong>Variable</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>poorhealth</td>
<td>Coded 1 if respondent has poor health, 0 otherwise</td>
</tr>
<tr>
<td>black</td>
<td>Coded 1 if black, 0 otherwise</td>
</tr>
<tr>
<td>rural</td>
<td>Coded 1 if respondent lives in a rural area, 0 otherwise</td>
</tr>
<tr>
<td>highlead</td>
<td>Coded 1 if the respondent has a high lead concentration in his or her body, 0 otherwise</td>
</tr>
</tbody>
</table>

They obtain the following results:

```
.nestreg, lr: logit poorhealth black rural highlead
```

**Block 1: black**

```
Iteration 0: log likelihood = -12353.146
Iteration 1: log likelihood = -12245.327
Iteration 2: log likelihood = -12229.778
Iteration 3: log likelihood = -12229.693
Iteration 4: log likelihood = -12229.693
```

Logistic regression

```
Number of obs = 49400  
LR chi2(1) = 246.91  
Prob > chi2 = 0.0000  
Pseudo R2 = 0.0100
```

Log likelihood = -12229.693

| poorhealth | Coef. | Std. Err. | z   | P>|z|  | [95% Conf. Interval] |
|------------|-------|-----------|-----|-----|----------------------|
| black      | 0.784 | 0.04665   | 16.80| 0.000| 0.6923975 - 0.8752686 |
| _cons      | -2.716 | 0.0197    | -137.95| 0.000| -2.7545 - 2.677668   |

**Block 2: rural**

```
Iteration 0: log likelihood = -12353.146
Iteration 1: log likelihood = -12079.44
Iteration 2: log likelihood = -12057.058
Iteration 3: log likelihood = -12056.937
Iteration 4: log likelihood = -12056.937
```

Logistic regression

```
Number of obs = 49400  
LR chi2(2) = 592.42
Prob > chi2 = 0.0000
Pseudo R2 = 0.0240
```

Log likelihood = -12056.937

| poorhealth | Coef. | Std. Err. | z   | P>|z|  | [95% Conf. Interval] |
|------------|-------|-----------|-----|-----|----------------------|
| black      | 1.008 | 0.0488    | 20.65| 0.000| 0.9120278 - 1.103297 |
| rural      | 0.695 | 0.0373    | 18.61| 0.000| 0.62196  - 0.7683478 |
| _cons      | -3.046 | 0.0285  | -106.91| 0.000| -3.101805 - 2.990124 |

Sociology 63993—Exam 3—Page 2
Block 3: highlead

Iteration 0:  log likelihood = [1]
Iteration 1:  log likelihood = -12072.72
Iteration 2:  log likelihood = -12049.04
Iteration 3:  log likelihood = -12049.03
Iteration 4:  log likelihood = -12049.03

Logistic regression                               Number of obs   =      49400
LR chi2(3)      =    [2]
Prob > chi2     =     0.0000
Log likelihood = -12048.91                       Pseudo R2       =     0.0246

-----------------------------------
poorhealth |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------
black |   .9879906   .0490653    20.00   0.000     .8918244    1.084157
rural |   .6990782   .0373575    18.71   0.000     .6258589    .7722975
highlead |   .2798416   .0677367     4.13   0.000     .1470801    .4126031
_cons |  -3.063828   .0288938  -106.04   0.000   -3.120449  -3.007197

------------------------------------------------------------------------------

| Block |        LL       LR     df  Pr > LR       AIC       BIC |
|-------|----------------|---------|--------|----------------|---------|
| 1     | -12229.69      246.91  1     0.0000   24463.38   24480.99 |
| 2     | -12056.94      345.50  1     0.0000   24119.87   24146.35 |
| 3     | -12048.91      16.06  1     0.0001   24105.82   24141.05 |

Based on the printout above, answer the following.

a. (6 points) Fill in the missing items [1], [2] and [3]. (HINT: You either need no calculations, or very easy ones.)

b. (7 points) The researchers decided that the last model (Block 3) was best. Explain why, and explain what this model tells you about the effects of race, residence and body lead on health.

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

<table>
<thead>
<tr>
<th>Race</th>
<th>Residence</th>
<th>Lead</th>
<th>Log odds</th>
<th>Odds</th>
<th>P(Poorhealth)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Rural</td>
<td>Low Lead</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Black</td>
<td>Non Rural</td>
<td>Low lead</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
d. (6 points) The researchers also ran the following:

```
.tab1 poorhealth if e(sample)
```

-

```
tabulation of poorhealth if e(sample)

RECODE of | health | (1=excellent | t,..., | 5=poor) | Freq. | Percent | Cum.

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
</table>
| 0    | 46,010       93.14       93.14
| 1    | 3,390        6.86       100.00

Total | 49,400      100.00
```

```
estat clas
```

- Logistic model for poorhealth

```
-------- True --------
Classified | D | ~D | Total
-----------+-------------------
+ | 0 | 0 | 0
- | 3390 | 46010 | 49400

Total | 3390 | 46010 | 49400
```

Classified + if predicted Pr(D) >= .5

True D defined as poorhealth != 0

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>Pr(+</td>
<td>D)</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>Pr( -</td>
<td>~D)</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>Pr( D</td>
<td>+)</td>
<td>.%</td>
<td></td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>Pr(~D</td>
<td>-)</td>
<td>93.14%</td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>False + rate for true ~D</td>
<td>Pr(+</td>
<td>~D)</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>False - rate for true D</td>
<td>Pr( -</td>
<td>D)</td>
<td>100.00%</td>
<td></td>
</tr>
<tr>
<td>False + rate for classified +</td>
<td>Pr(~D</td>
<td>+)</td>
<td>.%</td>
<td></td>
</tr>
<tr>
<td>False - rate for classified -</td>
<td>Pr( D</td>
<td>-)</td>
<td>6.86%</td>
<td></td>
</tr>
</tbody>
</table>

Correctly classified 93.14%

The medical researcher, who has little statistical background, was very impressed when he saw that 93.14% of the cases were correctly classified by the model. The sociologist, who has extensive training in statistics, was not very impressed. Explain why.

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 January 21, 2009. After Al Gore’s stunning decision to decline the nomination, Hillary Clinton successfully united her party and went on to win the Presidency in the largest Democratic landslide in history. But now, only hours after her inauguration, it is 3 a.m., the phone in the Oval Office is ringing – and Clinton must face the challenge she has spent a lifetime preparing for.
Clinton’s economic advisors have just learned that five of the largest banks in the country, along with mortgage lending giants Fannie Mae and Freddie Mac, are planning to declare bankruptcy at noon. Their poor decisions in the subprime loan market, combined with the disastrous consequences of the predatory and exploitive practices that some of them had engaged in, have left these institutions teetering on collapse. Clinton knows that many Americans will feel the government should do nothing to keep these companies afloat, and she shares the scorn for the corporate greed that has produced this crisis. At the same time, there is no doubt in her mind that inaction will result in millions of innocent people losing their homes, the further collapse of financial markets, and possibly even a depression the likes of which has not been seen since 1929.

She therefore is convinced that bold and decisive government action is needed. Anticipating this crisis, she prepared a plan months ago - but what is the best way to rally the American people to her side? By 6 a.m., she and her advisors have already crafted two radically different speeches in support of her plan – and by 8 a.m. 200 randomly selected Americans have been recruited for their feedback. Half of them will hear one of the speeches while the other half will hear the other. After hearing the talks, respondents will be asked to rate their support for the President’s plan on a five point scale, ranging from strongly agree to strongly disagree.

Armed with this knowledge, at 11 a.m. a calm, collected and determined Hillary Clinton will give the most important speech of her life before a worried nation.

b. President Bush is discouraged by polls that show he is one of the least popular presidents in recent American history. He thinks the American people are not aware of many of the good things he has done, such as his contributions to the fight against HIV/AIDS in Africa. He wants to know whether efforts to publicize these accomplishments would help his popularity. Two hundred randomly selected Americans will therefore be asked to rate Bush on a scale that ranges from 0 to 100. They will then see a film about Bush’s efforts in the war against AIDS. After the film, respondents will once again be asked to rate Bush.

c. The Democratic primary campaign has lasted far longer than anyone expected. It has taken a toll on the staff of both campaigns, with many quitting along the way while others persisted. For future reference, the Obama campaign wants to know the factors that affect how long someone stays with the campaign. It suspects that factors such as age, marital status, and past experience may all be important. For 300 randomly selected staffers, it has therefore collected information on (a) whether the staffer is still with the campaign (b) how long the staffer was or has been with the campaign (c) the marital status of the staffer when hired (d) whether the staffer had previously worked on a presidential campaign, and (e) the age of the staffer when hired.

d. A researcher believes that the following model describes her data. All variables are continuous.
e. A researcher has collected data on two scales – x (interest in politics) and y (participation in political activities). She is puzzled that the regression of y on x does not show as clear a relationship as she expected. To clarify what is going on, she creates the following graph:

![Graph showing relationship between x and y]

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

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

2. 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.