

Sociology 63993  
Exam 1  
February 13, 2014

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

1. Adjusted  $R^2$  is used to correct for heteroskedasticity.
2. Passive imputation should be used when one of the variables in the model is an interaction term that is the product of variables that have missing data.
3. Studentized residuals, Predictive Mean Matching (PMM), and Cook's distance are among the measures that can be used to detect outliers.
4. Pairwise deletion of missing data is commonly used when skip patterns have caused data for some cases to be missing.
5. If possible, increasing the sample size is one way to reduce problems of multicollinearity.

---

II. *Short answer.* Discuss all three of the following problems. (15 points each, 45 points total.) In each case, the researcher has used Stata to test for a possible problem, concluded that there is a problem, and then adopted a strategy to address that problem. Explain (a) what problem the researcher was testing for, and why she concluded that there was a problem, (b) the rationale behind the solution she chose, i.e. how does it try to address the problem, and (c) one alternative solution she could have tried, and why. (NOTE: a few sentences on each point will probably suffice – you don't have to repeat everything that was in the lecture notes.)

II-1.

. reg y x

| Source   | SS         | df   | MS         | Number of obs = 3975 |   |         |
|----------|------------|------|------------|----------------------|---|---------|
| Model    | 7588.90836 | 1    | 7588.90836 | F( 1, 3973)          | = | 0.33    |
| Residual | 92495852   | 3973 | 23281.1105 | Prob > F             | = | 0.5681  |
| Total    | 92503441   | 3974 | 23277.1618 | R-squared            | = | 0.0001  |
|          |            |      |            | Adj R-squared        | = | -0.0002 |
|          |            |      |            | Root MSE             | = | 152.58  |

| y     | Coef.     | Std. Err. | t     | P> t  | [95% Conf. Interval] |          |
|-------|-----------|-----------|-------|-------|----------------------|----------|
| x     | -1.606451 | 2.813712  | -0.57 | 0.568 | -7.122906            | 3.910004 |
| _cons | 2.176363  | 4.323281  | 0.50  | 0.615 | -6.299693            | 10.65242 |

```
. predict rstudent, rstudent
. extremes rstudent y x
```

| obs:  | rstudent  | y         | x         |
|-------|-----------|-----------|-----------|
| 3124. | -1920.873 | -9610.815 | 2.340901  |
| 2546. | -.1171118 | -14.99421 | -.4293181 |
| 3950. | -.1160397 | -15.22504 | -.1853033 |
| 3417. | -.1099706 | -14.32139 | -.1716456 |
| 2120. | -.108885  | -14.19914 | -.1448071 |

|       |          |          |          |
|-------|----------|----------|----------|
| 3082. | .1290129 | 18.32841 | 2.197414 |
| 3914. | .1295774 | 18.58599 | 2.09107  |
| 701.  | .1302823 | 18.6424  | 2.122781 |
| 1613. | .1351598 | 18.73533 | 2.526286 |
| 2008. | .1353822 | 19.60719 | 2.006944 |

```
. drop rstudent
. reg y x if y > -9600
```

| Source   | SS         | df   | MS         |
|----------|------------|------|------------|
| Model    | 10441.431  | 1    | 10441.431  |
| Residual | 99464.251  | 3972 | 25.0413522 |
| Total    | 109905.682 | 3973 | 27.6631467 |

```
Number of obs = 3974
F( 1, 3972) = 416.97
Prob > F = 0.0000
R-squared = 0.0950
Adj R-squared = 0.0948
Root MSE = 5.0041
```

| y     | Coef.    | Std. Err. | t     | P> t  | [95% Conf. Interval] |          |
|-------|----------|-----------|-------|-------|----------------------|----------|
| x     | 1.884699 | .0922977  | 20.42 | 0.000 | 1.703744             | 2.065655 |
| _cons | .1503674 | .1417923  | 1.06  | 0.289 | -.127625             | .4283598 |

```
. predict rstudent if e(sample), rstudent
(1 missing value generated)
. extremes rstudent y x
```

| obs:  | rstudent  | y         | x         |
|-------|-----------|-----------|-----------|
| 3885. | -3.361802 | -14.41269 | 1.186365  |
| 708.  | -3.228994 | -14.38133 | .8515745  |
| 2100. | -3.125682 | -10.08796 | 2.853061  |
| 1451. | -3.048306 | -11.82005 | 1.732555  |
| 3950. | -3.007255 | -15.22504 | -.1853033 |

|       |          |          |           |
|-------|----------|----------|-----------|
| 2606. | 3.123689 | 16.90548 | .6070455  |
| 3519. | 3.133599 | 18.47441 | 1.412693  |
| 2008. | 3.136441 | 19.60719 | 2.006944  |
| 1746. | 3.18355  | 15.73826 | -.1682539 |
| 619.  | 3.41969  | 18.23103 | .5278474  |

//-2.

. reg y x

| Source   | SS         | df  | MS         | Number of obs = 200 |   |         |
|----------|------------|-----|------------|---------------------|---|---------|
| Model    | 54113.6335 | 1   | 54113.6335 | F( 1, 198)          | = | 0.50    |
| Residual | 21545981.1 | 198 | 108818.086 | Prob > F            | = | 0.4815  |
| Total    | 21600094.7 | 199 | 108543.189 | R-squared           | = | 0.0025  |
|          |            |     |            | Adj R-squared       | = | -0.0025 |
|          |            |     |            | Root MSE            | = | 329.88  |

| y     | Coef.    | Std. Err. | t    | P> t  | [95% Conf. Interval] |          |
|-------|----------|-----------|------|-------|----------------------|----------|
| x     | 1.654202 | 2.345772  | 0.71 | 0.482 | -2.971701            | 6.280105 |
| _cons | 202.0068 | 23.36188  | 8.65 | 0.000 | 155.9367             | 248.0768 |

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of y

chi2(1) = 69.81

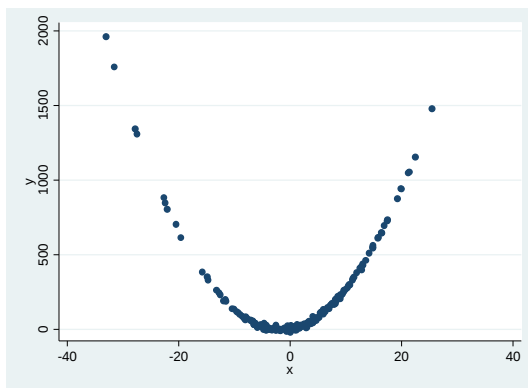
Prob > chi2 = 0.0000

. estat imtest

Cameron & Trivedi's decomposition of IM-test

| Source             | chi2   | df | p      |
|--------------------|--------|----|--------|
| Heteroskedasticity | 153.96 | 2  | 0.0000 |
| Skewness           | 30.99  | 1  | 0.0000 |
| Kurtosis           | 3.81   | 1  | 0.0510 |
| Total              | 188.76 | 4  | 0.0000 |

. scatter y x



. gen xsquare = x \* x

. reg y x xsquare

| Source   | SS         | df  | MS         | Number of obs = 200    |  |  |
|----------|------------|-----|------------|------------------------|--|--|
| Model    | 21580153.5 | 2   | 10790076.8 | F( 2, 197) = .         |  |  |
| Residual | 19941.1507 | 197 | 101.224115 | Prob > F = 0.0000      |  |  |
| Total    | 21600094.7 | 199 | 108543.189 | R-squared = 0.9991     |  |  |
|          |            |     |            | Adj R-squared = 0.9991 |  |  |
|          |            |     |            | Root MSE = 10.061      |  |  |

| y       | Coef.    | Std. Err. | t      | P> t  | [95% Conf. Interval] |          |
|---------|----------|-----------|--------|-------|----------------------|----------|
| x       | 6.893609 | .0724412  | 95.16  | 0.000 | 6.75075              | 7.036469 |
| xsquare | 1.993747 | .0043234  | 461.15 | 0.000 | 1.985221             | 2.002273 |
| _cons   | 1.356793 | .8348716  | 1.63   | 0.106 | -.28964              | 3.003225 |

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of y

chi2(1) = 1.22

Prob > chi2 = 0.2686

. estat imtest

Cameron & Trivedi's decomposition of IM-test

| Source             | chi2 | df | p      |
|--------------------|------|----|--------|
| Heteroskedasticity | 4.75 | 4  | 0.3140 |
| Skewness           | 1.57 | 2  | 0.4553 |
| Kurtosis           | 1.36 | 1  | 0.2435 |
| Total              | 7.68 | 7  | 0.3613 |

//-3.

. reg iq fatheriq motheriq

| Source   | SS         | df   | MS         | Number of obs = 8708   |  |  |
|----------|------------|------|------------|------------------------|--|--|
| Model    | 554164.753 | 2    | 277082.376 | F( 2, 8705) = 1636.98  |  |  |
| Residual | 1473444.77 | 8705 | 169.264189 | Prob > F = 0.0000      |  |  |
| Total    | 2027609.52 | 8707 | 232.871198 | R-squared = 0.2733     |  |  |
|          |            |      |            | Adj R-squared = 0.2731 |  |  |
|          |            |      |            | Root MSE = 13.01       |  |  |

| iq       | Coef.    | Std. Err. | t      | P> t  | [95% Conf. Interval] |           |
|----------|----------|-----------|--------|-------|----------------------|-----------|
| fatheriq | .654312  | .0299545  | 21.84  | 0.000 | .5955941             | .71303    |
| motheriq | .6981873 | .0148528  | 47.01  | 0.000 | .6690722             | .7273023  |
| _cons    | -34.6617 | 2.644898  | -13.11 | 0.000 | -39.84633            | -29.47708 |

```
. sum iq fatheriq motheriq
```

| Variable | Obs   | Mean      | Std. Dev. | Min       | Max    |
|----------|-------|-----------|-----------|-----------|--------|
| iq       | 10337 | 71.90088  | 15.35515  | 30.84     | 175.88 |
| fatheriq | 8708  | -16.54693 | 4.764332  | -62.33333 | -5     |
| motheriq | 10337 | 167.6512  | 9.660012  | 135.5     | 200    |

```
. gen one = 1
```

```
. impute fatheriq one, gen(fatheriq2)
```

15.76% (1629) observations imputed

```
. gen mdfatheriq = missing(fatheriq)
```

```
. sum iq fatheriq motheriq fatheriq2 mdfatheriq
```

| Variable   | Obs   | Mean      | Std. Dev. | Min       | Max    |
|------------|-------|-----------|-----------|-----------|--------|
| iq         | 10337 | 71.90088  | 15.35515  | 30.84     | 175.88 |
| fatheriq   | 8708  | -16.54693 | 4.764332  | -62.33333 | -5     |
| motheriq   | 10337 | 167.6512  | 9.660012  | 135.5     | 200    |
| fatheriq2  | 10337 | -16.54693 | 4.372804  | -62.33333 | -5     |
| mdfatheriq | 10337 | .1575892  | .364373   | 0         | 1      |

```
. reg iq fatheriq2 motheriq mdfatheriq
```

| Source   | SS         | df    | MS         | Number of obs = | 10337   |
|----------|------------|-------|------------|-----------------|---------|
| Model    | 643417.937 | 3     | 214472.646 | F( 3, 10333) =  | 1235.58 |
| Residual | 1793609.31 | 10333 | 173.580694 | Prob > F =      | 0.0000  |
|          |            |       |            | R-squared =     | 0.2640  |
|          |            |       |            | Adj R-squared = | 0.2638  |
| Total    | 2437027.25 | 10336 | 235.7805   | Root MSE =      | 13.175  |

| iq         | Coef.    | Std. Err. | t      | P> t  | [95% Conf. Interval] |
|------------|----------|-----------|--------|-------|----------------------|
| fatheriq2  | .6524215 | .0302144  | 21.59  | 0.000 | .5931955 .7116475    |
| motheriq   | .7025812 | .0136778  | 51.37  | 0.000 | .67577 .7293923      |
| mdfatheriq | 2.142615 | .3556721  | 6.02   | 0.000 | 1.445429 2.839801    |
| _cons      | -35.4298 | 2.444892  | -14.49 | 0.000 | -40.22227 -30.63734  |

*III. Computation and interpretation.* (35 points total) A recent unpublished study by Princeton researchers claims that Facebook will lose 80% of its users by 2015-2017. Although Facebook claims that the study is ridiculous, it is concerned about its long-term popularity and growth. It has therefore collected data on the following variables from a random sample of 7,236 active Facebook members.

| Variable | Description                                                                                                                                        |
|----------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| stayfb   | Scale that measures the likelihood the respondent will still be a Facebook user in five years. Potentially ranges from a low of 0 to a high of 100 |
| frdties  | Strength of the respondent's current friendship ties on Facebook. Potentially ranges from a low of 0 to a high of 100                              |
| age      | Age in years                                                                                                                                       |
| female   | 1 = female, 0 = male                                                                                                                               |
| white    | 1 = white, 0 = not white                                                                                                                           |

An analysis of the data yields the following results. [NOTE: You'll need some parts of the following to answer the questions, but other parts are extraneous. You'll have to figure out which is which.]

```
. reg stayfb frdties age female white, beta
```

| Source   | SS         | df   | MS         | Number of obs = | 7236   |
|----------|------------|------|------------|-----------------|--------|
| Model    | 70915.1394 | [2]  | 17728.7848 | F( 4, 7231) =   | [1]    |
| Residual | 204149.201 | 7231 | 28.2324992 | Prob > F =      | 0.0000 |
| Total    | 275064.341 | 7235 | 38.0185682 | R-squared =     | [3]    |
|          |            |      |            | Adj R-squared = | 0.2574 |
|          |            |      |            | Root MSE =      | 5.3134 |

| stayfb  | Coef.     | Std. Err. | t      | P> t  | Beta      |
|---------|-----------|-----------|--------|-------|-----------|
| frdties | .2951861  | .009427   | [4]    | 0.000 | .4638028  |
| age     | -.0458587 | .0037968  | -12.08 | 0.000 | -.1278213 |
| female  | .7808451  | .1787656  | 4.37   | 0.000 | .0632679  |
| white   | 1.641666  | .2021571  | 8.12   | 0.000 | .0823415  |
| _cons   | 67.02458  | .3835846  | 174.73 | 0.000 | .         |

```
. vif
```

| Variable | VIF  | 1/VIF    |
|----------|------|----------|
| frdties  | 2.14 | 0.467839 |
| female   | 2.04 | [5]      |
| age      | 1.09 | 0.916446 |
| white    | 1.00 | 0.998317 |
| Mean VIF | 1.57 |          |

```
. test frdties age female white
```

```
( 1) frdties = 0
( 2) age = 0
( 3) female = 0
( 4) white = 0
```

```
F( 4, 7231) = 627.96
Prob > F = 0.0000
```

```
. sum
```

| Variable | Obs  | Mean     | Std. Dev. | Min    | Max    |
|----------|------|----------|-----------|--------|--------|
| stayfb   | 7236 | 79.21281 | 6.16592   | 37.648 | 94.028 |
| frdties  | 7236 | 42.35774 | 9.688028  | 10     | 74.5   |
| age      | 7236 | 47.70854 | 17.18618  | 20     | 74     |
| female   | 7236 | .5210061 | .4995931  | 0      | 1      |
| white    | 7236 | .8928966 | .309266   | 0      | 1      |
| black    | 7236 | .1071034 | .309266   | 0      | 1      |

```
. test white = female
```

```
( 1) - female + white = 0
```

```
      F( 1, 7231) =    10.23
      Prob > F =    0.0014
```

```
. estat hettest
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of stayfb

```
      chi2(1)      =    0.55
      Prob > chi2   =    0.4591
```

```
. estat imtest, white
```

White's test for Ho: homoskedasticity  
against Ha: unrestricted heteroskedasticity

```
      chi2(12)     =   107.11
      Prob > chi2   =    0.0000
```

Cameron & Trivedi's decomposition of IM-test

| Source             | chi2   | df | p      |
|--------------------|--------|----|--------|
| Heteroskedasticity | 107.11 | 12 | 0.0000 |
| Skewness           | 178.07 | 4  | 0.0000 |
| Kurtosis           | 27.67  | 1  | 0.0000 |
| Total              | 312.86 | 17 | 0.0000 |

```
. pcorr2 stayfb frdties age female white
```

```
(obs=7236)
```

Partial and Semipartial correlations of stayfb with

| Variable | Partial | SemiP   | Partial^2 | SemiP^2 | Sig.  |
|----------|---------|---------|-----------|---------|-------|
| frdties  | 0.3456  | 0.3172  | 0.1194    | 0.1006  | 0.000 |
| age      | -0.1406 | -0.1224 | 0.0198    | 0.0150  | 0.000 |
| female   | 0.0513  | 0.0443  | 0.0026    | 0.0020  | 0.000 |
| white    | 0.0951  | 0.0823  | 0.0090    | 0.0068  | 0.000 |

```
. alpha frdties age female white, s
```

Test scale = mean(standardized items)

```
Average interitem correlation:    0.1590
Number of items in the scale:      4
Scale reliability coefficient:      0.4307
```

a) (10 pts) Fill in the missing quantities [1] – [5]. (A few other values may have also been blanked out, but you don't need to fill them in.)

b) (25 points) Answer the following questions about the analysis and the results, explaining how the printout supports your conclusions.

1. Summarize the key findings. Which groups or types of individuals are most likely to be on Facebook 5 years from now and which are least likely?

2. The researchers originally planned to use multiple imputation to deal with missing data, but then it turned out there was no missing data (Facebook knows everything about you). If there had been missing data on frdties and female, what imputation method or methods would you recommend using (e.g. logit, mlogit, regress, ologit, pmm, poisson, or something else)? Briefly explain why.

3. The researchers were concerned that the items may suffer from random measurement error. Would you encourage them to create a scale out of the independent variables in order to deal with the problem?

4. The Breusch-Pagan / Cook-Weisberg test for heteroskedasticity is insignificant. Would you therefore agree that there is no need to be worried about heteroskedasticity?

5. Facebook wants to know which variable has the greatest impact on whether or not somebody is likely to stay with Facebook. What would you say? Cite at least two or three results from the printout to support your argument.