midterm

This version: March 20, 2018

Name: ..... **Blue Book #:.....**

- put your name and blue book # above; put ONLY the number on the blue book; make up the number–eg last 4 digits of your phone #, your license plate # (have at least 3 digits, not 123, or 001, etc–needs to be unique!)
- the exam is open book, open note, and calculators are allowed (all other electronic devices are prohibited)
- per calculations: show your full work and be sure the logic of your calculations is clear; in most cases, a "naked number" won't do; and always interpret your results!
- be clear, concise, to the point! don't dump everything you know on all related topics and ask me to pick what is right and discard what is wrong!
- write legibly–if i cannot read it, i cannot grade it!
- each subquestion is worth the same credit (if no subquestions, whole question is worth as much as a subquestion)
- hand in BOTH your test and blue book, in SEPARATE piles

1 The effect of non-perfect collinearity on estimates is really the same as the effect of small sample. Do you agree or disagree and why?

yes, both blow up std err!

2 Imagine, you are doing community development! And you have data on number of meetings with local people and data on voting turnout at neighborhood level, and you think that the more community meetings, the higher the proportion of neighborhood votes in elections. But as number of meetings increases, the voting proportion increases at a decreasing rate. In other words, increase from very few to few community meetings has large positive effect on voting, but an increase from some to many has little effect on voting. How would you model such relationship? Be specific, write down the model and describe it.

quadratic with + and - on sq; or one of the log models with decreasing returns

3 Briefly (in few sentences) tell me why would you prefer multivariate regression over bivariate correlation?

multiple controls!

4 You have the following data for 3 people (in hundreds of pounds and hundreds of thousands of dollars):

- Kate: weight:1 lbs; salary:$2
- Amy: weight: 1.5 lbs; salary:$1
- Bob: weight: 3 lbs; salary:$.5

[NOTE: you may observe: "Nonsensical to use a sample size of 3 to do a regression" Right, yes; still please calculate–the idea is to test that you can calculate; and cannot have like 30 obs–then it would take hours to calculate]

a) Using regression test a hypothesis that heavier people make less money. [don't forget about statistical significance!]
b) Which of the 3 persons had worse prediction, i.e. in which case your estimate of salary was off by largest amount?

regression; we do not know height! for obesity variable

```
input wei sal
1 2
1.5 1
3 .5
end
```

reg sal wei

```
      Source |       SS           df       MS      Number of obs   =         3
-------------+----------------------------------   F(1, 1)         =      3.85
       Model |  .926282051         1  .926282051   Prob > F        =    0.3000
    Residual |  .240384615         1  .240384615   R-squared       =    0.7940
-------------+----------------------------------   Adj R-squared   =    0.5879
       Total |  1.16666667         2  .583333333   Root MSE        =    .49029

------------------------------------------------------------------------------
         sal |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wei |  -.6538462   .3330867    -1.96   0.300    -4.886114    3.578422
       _cons |   2.365385   .6730769     3.51   0.176    -6.186869    10.91764
------------------------------------------------------------------------------
```

make sure you interpret right in terms of hundreds of thousands!

predict r,resid

l

```
     +----------------------+
     | wei   sal          r |
     |----------------------|
  1. |   1     2   .2884615 |
  2. | 1.5     1  -.3846154 |
  3. |   3    .5   .0961538 |
     +----------------------+
```

d) pos bias

5 A young scholar from Rowan University is interested in income disparities by race. She hypothesizes that blacks make less than whites, even controlling for education. She has income variable measured in $, education variable measured as years of schooling (educ). And she has race of household variable (hhrace) measured in a following way:

```
 tabulation:  Freq.   Numeric  Label
             44,482         1  white
              7,662         2  black
                280         3  amer indian
                822         4  asiatic, oriental
              2,294         5  other, mixed
```

She estimates a regression with following results:

```
. reg inc hhrace educ
reg inc hhrace educ

      Source |       SS           df       MS      Number of obs   =    45,769
-------------+----------------------------------   F(2, 45766)     =   4497.13
       Model |  14934.3297         2  7467.16485   Prob > F        =    0.0000
    Residual |  75991.1246    45,766  1.66042749   R-squared       =    0.1642
-------------+----------------------------------   Adj R-squared   =    0.1642
       Total |  90925.4543    45,768  1.98665999   Root MSE        =    1.2886

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hhrace |      -0.11       0.01   -15.65    0.00        -0.12       -0.09
        educ |       0.18       0.00    93.13    0.00         0.18        0.18
       _cons |       0.68       0.03    24.42    0.00         0.62        0.73
------------------------------------------------------------------------------
```

a) Do the results support her hypothesis? Why or why not?
b) How would you improve her model?

6 Professor X is analyzing the effect of various socio-demographic measures on wellbeing indicators across US counties. Variables are defined in table 1. And results are set in table 2. [As it is usually the practice in published research, dependent variables are specified as column headers and independent variables are specified in each row, and slope coefficients are in the body of the table].

a) Is there support for hypothesis that smoking results in lower wellbeing? Why?
b) Which variable is the strongest predictor of wellbeing? Why?

c) Clearly unemployment and lack of insurance are highly correlated, so there is multicollinearity! But the theory says that both do predict wellbeing (ie, both belong in the model). What should we do? Drop a collinear variable or keep it? Why? a) yes; crim is positive
b) can look at stats significance; but otherwhise without beta coefficients or sum stats showing ranges and possibly std cannot really tell
c)

Table 1: Variable definitions

name	description
mentally unhealthy days	"average number of reported mentally unhealthy days per month, for adults, Behavioral Risk Factor Surveillance System (BRFSS), 2002-2008"
physically unhealthy days	"average number of reported physically unhealthy days per month for adults, Behavioral Risk Factor Surveillance System (BRFSS), 2002-2008"
years lost	"age-adjusted years of potential life lost (YPLL) rate per 1000 persons, 2004-2006"
% low birthweight	"percent of births with low birth weight (<2500g),2000-2006"
sprawl index/100	"Ewing's sprawl index"
% obese	"percent of adults that report BMI >= 30, 06-08; National Center for Chronic Disease Prevention and Health Promotion"
gini	"gini coefficient, decennial census, 2000"
persistent poverty	"20 percent or more of residents were poor as measured by each of the last 4 censuses, 1970, 1980, 1990, and 2000"
ERS rural-urban	"2003 ERS Rural-Urban Continuum Code"
per capita income	"per capita personal income (USD 1,000), 2005"
no social-emotional support	"percent of adults that report not getting social/emotional support (2005-2008);BRFSS"
crime rate	"Index crime rate (per 100,000 persons), 2004"
% smokers	"Percent of adults that report smoking at least 100 cigarettes and that they currently smoke, Behavioral Risk Factor Surveillance System (BRFSS) 2002-2008"
% uninsured	"percent of adults 18-64 without insurance, Census/Current Population Survey (CPS) Small Area Health Insurance Estimates (SAHIE), 2005 "
% college	"percent of population age 25+ with 4-year college degree or higher, American Community Survey (ACS), 2005-2007"
% unemployed	"percent of population age 16+ unemployed and looking for work, Local Area Unemployment Statistics, Bureau of Labor Statistics, 2008"
%> 65	"percent population over 65, 2005"
% black	"percent black, 2005"

Table 2: OLS regressions of various measures of wellbeing.

	mentally unhealthy days	physically unhealthy days	years lost	% low birthweight
sprawl index/100	0.42***	0.45***	-0.14	0.09
no social-emotional support	0.04***	0.03***	0.06	0.00
crime rate	0.00**	-0.00	0.00***	0.00*
% obese	0.02*	0.01	0.59***	0.00
% uninsured	-0.02***	-0.02**	-0.47***	-0.05***
% college	0.00	-0.02***	-0.64***	-0.02***
% unemployed	0.02	0.02	-0.13	-0.07**
persistent poverty	-0.03	-0.01	-3.75	0.32
%> 65	0.02*	0.00	0.77***	0.02*
gini	0.01	0.02*	1.05***	0.08***
% black	-0.02***	-0.01***	0.38***	0.07***
ERS rural-urban	-0.01	-0.01	-0.52	-0.00
per capita income	-0.02***	-0.02***	-0.15	-0.01
% smokers	0.04***	0.03***	0.75***	0.03**
constant	1.78**	2.77***	18.50	5.22***
N	769	769	769	769

*** p<0.001, ** p<0.01, * p<0.05; robust std err