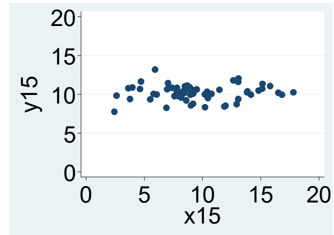
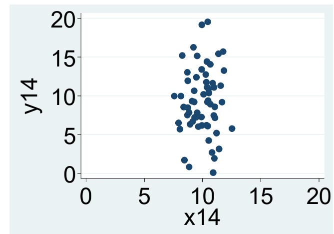
Homework 4: ECO220Y

Required Exercises: Chapter 6: 5, 9, 13, 15, 17, 20, 27 – 29, 31, 35 – 37*, 43 (*Note: For Exercise 37, you may find it convenient to use Excel to assist with the calculations.)

Required Problems:

(1) Answer and EXPLAIN your answer. How do the graphs to the right *differ*?





(A) the graph on the top shows no relationship
(B) the graph on the bottom shows a steeper relationship
(C) the graph on the bottom shows a stronger relationship
(D) the graph on the top shows a y variable with less variance
(E) All of the above

(2) Are each of these observational or experimental data? Explain and ***specifically apply*** course concepts to each.

(a) Data on the interest rate and growth rate of GDP over time in Canada

(b) Obesity rate in mice fed low-carb diet versus regular diet

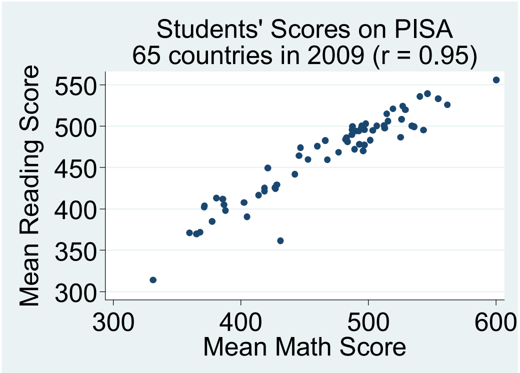
(c) Prices and quantities sold of bath tissue for randomly selected retail outlets

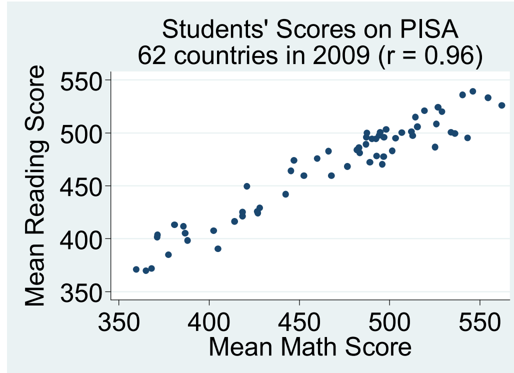
(3) Read Exercise 40 (in Chapter 6) in the textbook for background. To start, consider 2009 data. (You can view the 2009 data at: http://www.oecd.org/pisa/pisaproducts/48852548.pdf by looking at page 15 and clicking the StatLink at the bottom of the table.) Here is the variance-covariance matrix for those 2009 data.

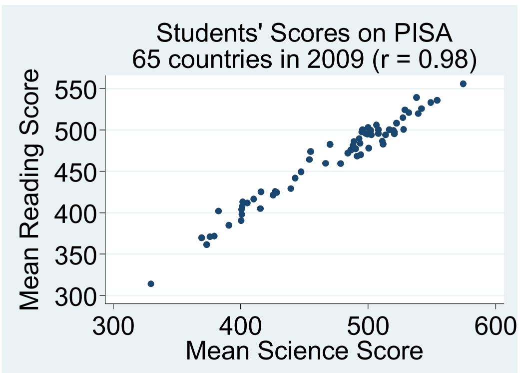
```
. corr reading math science, covariance;
(obs=65)
             |  reading     math  science
-------------+---------------------------
     reading |   2664.7
        math |  2926.14  3576.25
     science |  2841.29   3255.2  3143.86
```

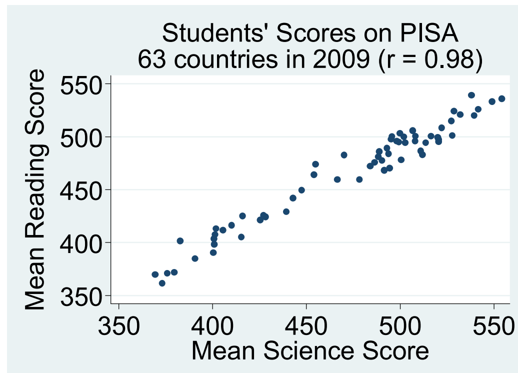
(a) Create a correlation matrix. (Note: You need only the information in the matrix above.)

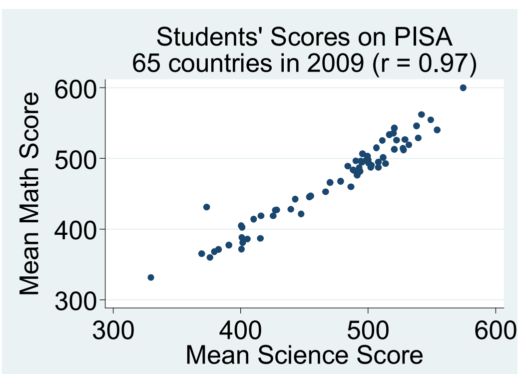
(b) Given the graphs below (next page), should we say that reading scores and math scores are strongly *correlated* or *associated*? Do any serious concerns about underlying conditions prevent us from saying either of these?

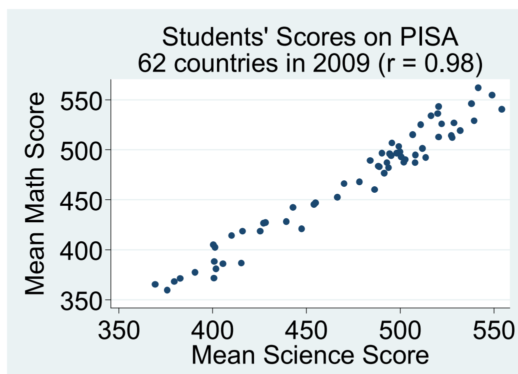












(c) Which kind of data are these? How does that affect the interpretation of the correlations?

(d) Consider the 2012 PISA data http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf (see page 19 and click the StatLink at the bottom of the table). Compare and contrast these results with those from 2009 discussed earlier.

```
. corr reading math science, covariance;
(obs=65)

             |   reading      math   science
-------------+------------------------------
     reading |   2216.65
        math |   2505.97   3075.51
     science |   2338.37   2736.36   2576.39
```

(4) Make sure you did not miss the required reading SW11 (available in electronic format on the Readings page of our Quercus site). Here is a multiple-choice quiz to help you assess your understanding of this reading.

(4.1) What does the phrase “empirical evidence” mean? It refers to evidence based on ____.

(A) many experts’ intuitions
(B) observations recorded as data
(C) the outcome of deductive reasoning
(D) the results of simulations derived from theory
(E) rigorous mathematical modeling of a phenomenon

► **Questions (4.2) – (4.4):** Consider the example about applying fertilizer to some randomly selected plots of farmland. Suppose there are 50 plots in the treatment group and 50 plots in the control group.

(4.2) This is an example of what kind of data?

(A) time series data
(B) experimental data
(C) observational data
(D) natural experiment data
(E) longitutudinal (panel) data

(4.3) What happens to the 50 plots in the control group?

(A) all get fertilizer
(B) none get fertilizer
(C) some selected plots get fertilizer
(D) some randomly selected plots get fertilizer
(E) all are subject to careful control such that they each receive exactly the same water, sunlight, weeding, seeds, wind, slope, etc.

(4.4) What is the distinguishing feature of a randomized controlled experiment in the farming example?

(A) that the 100 plots are randomly divided into the two groups
(B) that the 50 plots of land in each group are perfectly identical in every respect
(C) that the plots in the treatment group are carefully matched to plots in the control group
(D) that the randomization process has been controlled to ensure that the plots are comparable
(E) that the researchers have verified that all other variables are held constant across these plots

► **Questions (4.5) – (4.9):** Which could correctly fill in the blank. Data on the unemployment rate, inflation rate, and growth rate, in each province for each of the last twelve months would be an example of ____ data?
(A) correct **(B)** incorrect

(4.5) observational
(4.6) experimental
(4.7) cross-sectional
(4.8) time series
(4.9) longitudinal (panel)

(4.10) By using observational data that shows that in schools with smaller class sizes the learning outcomes are typically better than in other schools with larger classes, why is it difficult to answer “Does reducing class size improve elementary school education?” It is difficult because ____.

(A) there are no data available that quantitatively measure outcomes for learning
(B) there is a lot of variability in outcomes across students: each student is different

(C) class sizes vary little across schools making it hard to separate the signal from the noise
(D) factors like neighborhood wealth vary across schools and affect class sizes and outcomes
(E) all of the above

► **Questions (4.11) – (4.14):** You wonder if the format of a questionnaire affects how students respond to questions about their undergraduate experience. You select a random sample of 30 students and randomly divide them into two groups. One group answers the questionnaire online while the other does so using pen and paper. Which are valid criticisms of your study design? **(A)** valid; **(B)** not valid

(4.11) While the results will not by systematically wrong, there will be a fair bit of sampling noise and this will limit your ability to answer your research question.

(4.12) You have made no attempt to ensure that the two groups are *otherwise identical* and this means that your data should not be used to answer your research question.

(4.13) You have failed to ensure that other factors are *held constant* across the two groups and this will lead to an overestimate of the causal effect of the questionnaire format.

(4.14) You should have conducted a randomized controlled experiment rather than relying on observational data.

(5) Get comfortable using the terms "endogenous," "exogenous," and "endogeneity bias." Go back through the examples in Lecture 4 – Canadian inflation and interest rates, chocolate consumption and Nobel Laureate production, and drug dosage and hours of sleep – and apply these terms appropriately. Answer with several sentences for each case.

(6) Consider the cross-tabulation below of two dummy variables from a survey of 774 respondents in a study like Carlin et al. (2017) from Module A in DACM. The variable male takes a value of 1 if the respondent is male and 0 otherwise. The variable chosedom takes a value of 1 if the respondent chose the dominant credit card and 0 otherwise. What is the coefficient of correlation between these two variables?

```
           |         male
  chosedom |         0          1 |     Total
-----------+----------------------+----------
         0 |       178        215 |       393
         1 |       183        198 |       381
-----------+----------------------+----------
     Total |       361        413 |       774
```

(7) Consider the data and scatter diagram below (next page) from the Council of Ontario Universities. (Data retrieved from http://cou.on.ca/numbers/multi-year-data/enrolment/ on September 22, 2017.) It provides information on the number of full time equivalent (FTE) students enrolled in undergraduate (UG) programs across all of Ontario's universities annually since 2000.

(a) The coefficient of correlation between the variables year and UG_tot_FTEs is 0.9683. In light of the given background information, what does that value of the correlation mean?

(b) How would the coefficient of correlation change if enrolments were measured in 1,000s of FTE students?

(c) How would the coefficient of correlation change if year were recorded as 0, 1, 2, …, 16 instead of 2000, 2001, …, 2016?

(d) How would the coefficient of correlation change if the level of enrolment in the year 2000 were 420,000 instead of 244,945? Also, how would that affect its ability to summarize the strength of the relationship?

year	UG_tot_FTEs
2000	244945
2001	257488
2002	278765
2003	311660
2004	327371
2005	341882
2006	350030
2007	348611
2008	352945
2009	367901
2010	381583
2011	391502
2012	400272
2013	406407
2014	410086
2015	413206
2016	420687