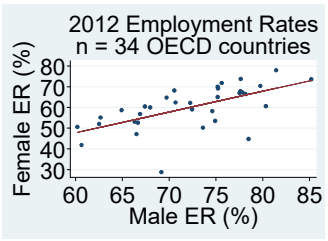
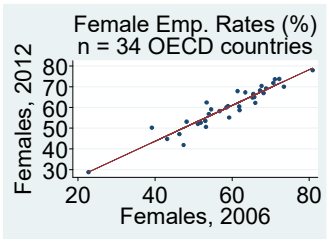
Homework 5: ECO220Y

Required Exercises: Chapter 7: 3, 5, 7, 8, 17, 44

Required Problems:

(1) Consider these two scatter plots and regression lines for the 34 member nations of the Organization for Economic Cooperation and Development (OECD). These summarize the relationship between the 2012 female employment rate (y variable) and either the 2012 male employment rate (x variable) or the 2006 female employment rate (x variable). (*Source:* OECD.Stat (https://stats.oecd.org/); Retrieved May 22, 2013.) ***Notice that the y variable is exactly the same in both***: this is crucial information when answering the below.





(a) Which has a steeper slope? Explain. (Be careful.)

(b) Which has a bigger intercept? Explain. (Be careful.)

(c) Which has a bigger mean of the residuals? Explain.

(d) Which has a bigger value of the s_e? Explain.

(e) Which has a bigger value of the SST? Explain.

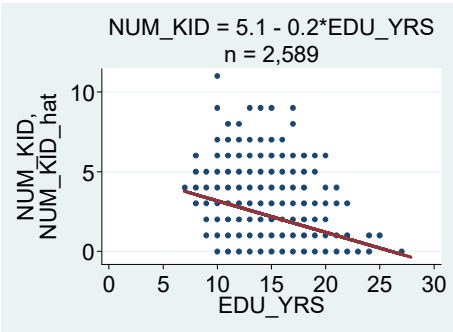
(f) Which has a bigger value of the R^2? Explain.

(g) Which has a bigger value of the SSR? Explain.

(h) Which has a bigger value of the SSE? Explain.

(i) Comparing and contrasting the two scatter plots, overall, which conclusion should we draw? (Answer briefly.)

(2) An economist studies the relationship between a female education (EDU_YRS) and number of children (NUM_KID). A sample of 2,589 Canadian women yields the following OLS line and coefficient of correlation:



NUM_KID-hat = 5.1 – 0.2*EDU_YRS
r = -0.3132

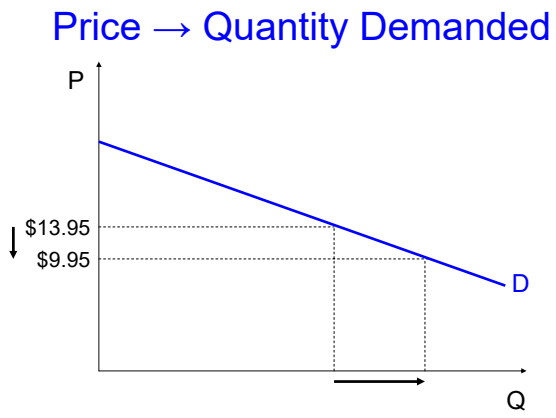
(a) Which kind of data are these?

(b) How should you interpret these results? What is the meaning of 5.1? What is the meaning of -0.2? What is the meaning of -0.3132?

(c) Is the scatter diagram a good description of these data? Explain. Aside statistics such as the correlation and least squares line, what is another way to summarize the relationship between these two variables?

(3) An important problem in economics is how to estimate demand. When analyzing a market – regardless of the model of competition – we specify demand. For example, it could be P = 100 – Q, which is a linear demand curve. The more

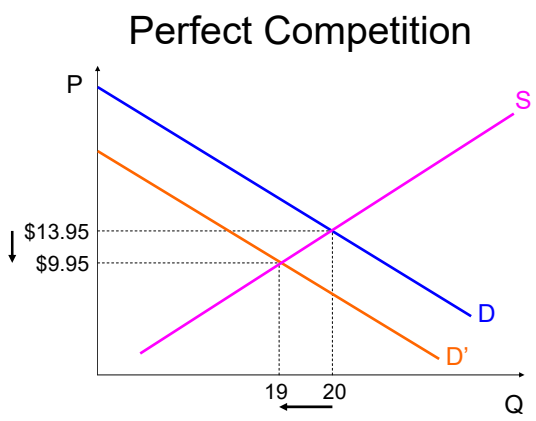
general linear form is P = a – bQ, where a and b are parameters that we estimate using data. Price affects quantity demanded. One of the undisputed tenets of economics is the Law of Demand: quantity demanded goes down as price goes up and vice versa.

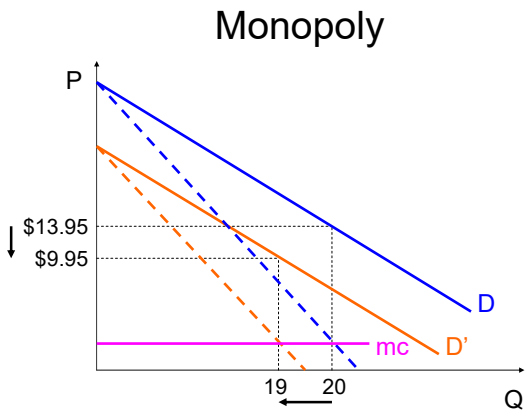


(a) Give an example of time series data containing prices and quantities of some good. Give an example of cross-sectional data containing prices and quantities of some good.

(b) When considering collecting market data on prices and quantities, which kind of data will typically be available: observational or experimental? Why?

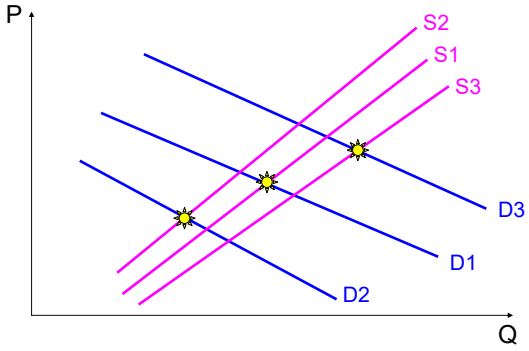
Estimating demand from observational data containing only prices and quantities is impossible. This is because demand shifters affect *both* the price and the quantity sold. This makes price an endogenous variable and means that there will be a serious endogeneity bias in the estimate of the slope of demand (i.e. for a simple linear demand) obtained from a regression analysis (least squares method). To see that demand shifters – population size, tastes, income, prices of substitute goods, prices of compliment goods, etc. – affect *both* the market price and the market quantity consider two simple models: perfect competition and monopoly.





(c) Redraw the diagram for the experimental drug trial data that we did as an example in lecture: that is, a diagram with boxes and arrows that illustrates the research question and the role of any other variables. Next draw such a diagram for the question of estimating demand where we ask how price affects the quantity demanded. Make sure to include the role of demand shifters in your diagram.

Henry J. Moore, "father of economic statistics," conduced regressions for many industries in early 20th century. In some regressions found negative demand elasticities, but in pig iron, for example, found positive demand elasticity and concluded "he had discovered a new type of demand curve with positive slope." Of course this is insane: I think we'd all like start businesses selling products where the higher the price we charge – other things equal (including quality, service, etc.) – the more our customers want to buy! Moore's regression analysis had a serious endogeneity bias and hence did not reflect the real slope of demand for pig iron at all. To illustrate consider the following diagram.

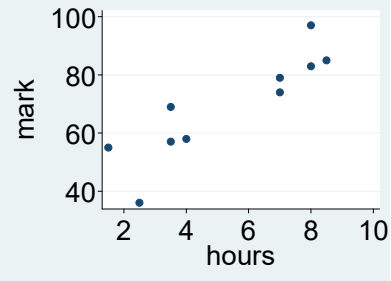


(d) If the above diagram shows three different time periods which kind of data would it generate on prices and quantities? (observational or experimental; time series, cross-sectional, or panel) How many variables? How many observations? Would the correlation be positive or negative? Does this mean demand is upward sloping?

(e) Suppose we had a random sample of 10,000 customers. At random each was offered the opportunity to buy a product at a price of $1, $1.50, $2, $2.5 or $3. We used their purchase decisions to construct data with 5 observations and two variables: price and total quantity sold. (Note that while the original data has 10,000 observations, the data on which we estimate demand will have only five data points: one for each unique value of the price variable. That is because demand tells the total quantity sold, which is summed over all customers, given the price.) Could we use these data to estimate the slope of demand? Would that estimate suffer from an endogeneity bias? Is it a problem that the 10,000 customers have various different tastes, incomes, etc.?

(4) How does time spent sleeping the night before a test affect a student's mark? For a random sample of 10 students in a large economics course, right after the test, each student is asked time slept the night before (in hours). Later the marks – points earned out of 120 points possible – on the test are recorded. Use a handheld calculator for all parts.

hours	mark
7	79
1.5	55
8.5	85
4	58
3.5	57
7	74
8	83
8	97
3.5	69
2.5	36



(a) Compute the covariance and indicate the units of measurement.

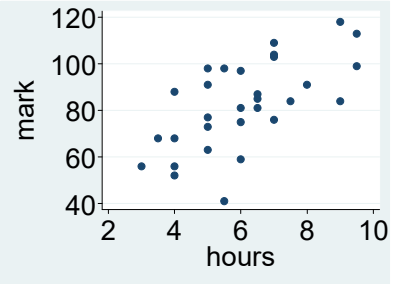
(b) Using your answer from (a) and the fact that the s.d. of sleep is 2.60395 hours and the s.d. of marks is 17.981781, compute the coefficient of correlation and interpret it.

(c) Compute coefficient of determination and interpret it.

(d) What kind of data are these: observational, experimental, or a natural experiment?

(e) Find the regression line equation and interpret the coefficients. Explain how you can or cannot use these results to answer the research question posed at the start of this problem.

(f) Consider trying to replicate the results with a fresh random sample (n = 30). Results are below and to the right. Presuming that both the original study and the replication study contained no non-sampling errors, what is the best explanation for each of the differences (i.e. the difference in the covariance, standard deviations, correlation, R^2, intercept, and slope)?



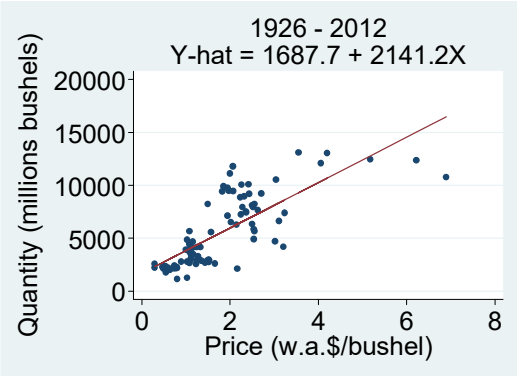
covariance = 21.750000; s.d. hours = 1.762068; s.d. marks = 19.098204; correlation = 0.64631462

OLS results: mark-hat = 39.8857 + 7.0050902*hours, n = 30, R-squared = 0.4177

(g) Consider again the results from Part (f). Suppose that marks are calculated as percentage marks (i.e. 0 – 100%) instead of raw marks out of 120 points. What would the new OLS results be? The new R-squared?

(h) Consider again the results from Part (f). Suppose that study time is recorded in minutes instead of hours. What would the new OLS results be? The new R-squared?

(5) The figure shows a regression analysis like Slide 10 of Lecture 5 except that it uses U.S. corn production instead of Manitoba corn production. Production in millions of bushels is the dependent variable and the weighted-average farm price in dollars per bushel is the independent variable. (Source of these data: https://www.ers.usda.gov/data-products/feed-grains-database/feed-grains-yearbook-tables/.)



(a) Is the OLS line an estimate of the supply curve?

(b) Interpret the OLS line.

(6) Consider the following excerpt from Daniel Kahneman's biography from the Nobel prize website.

> I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech, which began by conceding that positive reinforcement might be good for the birds, but went on to deny that it was optimal for flight cadets. He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case." This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback. We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa. But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency. http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2002/kahneman-bio.html

Consider Prof. Murdock's general explanation of regression to the mean:

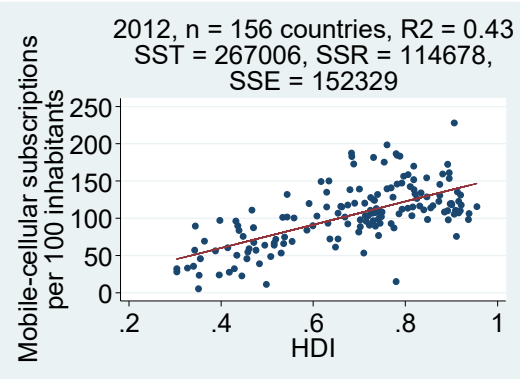
> Suppose observed performance is a function of some non-random factors (such as skill, effort, etc.) and a random component (pure chance). High draws of the random component boost performance and low draws detract from performance. However, if you get lucky with a good draw of the random component you should not expect to be lucky next time. You should expect an average draw, which if you had been lucky is worse than your lucky draw. Similarly, if you had a poor draw of the random component you should not expect to be unlucky next time: you should expect an average draw, which, of course, is better than poor. Hence there is regression towards the mean: people who got lucky tend to move down towards the mean and people who got unlucky tend to move up towards the mean. The more that performance is driven by random factors the bigger the regression to the mean effect will be. In the extreme case of Kahneman's example of tossing a coin behind your back two times not even knowing where the target is there is no skill and all chance involved: this was a great way to illustrate regression to the mean because it is extremely powerful in this case. On the other hand, if there is no random component to performance then there will be no regression to the mean.

(a) In what way was the seasoned flight instructor mistaken? (Is it that his recollection is faulty? Is it that his inference based on his observations is faulty? Is it both? Is it something else entirely?)

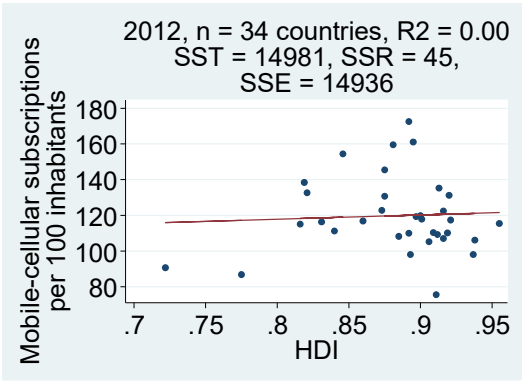
(b) Is it important for Kahneman's argument that some portion of a flight cadet's performance is entirely random and beyond his/her control?

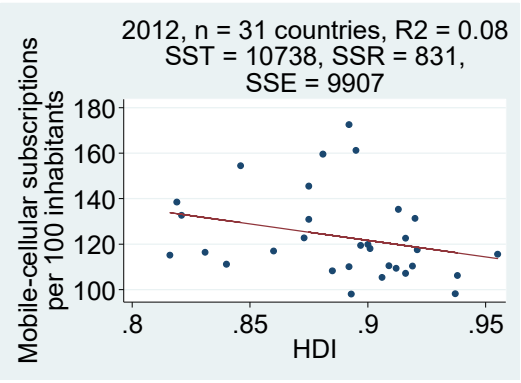
(c) What does the last line of Kahneman's speech mean for you? (i.e. Is he optimistic that we will be able to take the message of the simulation we just did to heart?)

(7) The United Nations (UN) reports the Human Development Index (HDI) across countries and over time. It is a "way of measuring development by combining indicators of life expectancy, educational attainment and income into a composite human development index, the HDI" (see "Human Development Index (HDI)" at http://hdr.undp.org/en/statistics/hdi). Consider cell phone penetration versus the HDI. The 2012 HDI data are downloaded from the UN website for a cross-section of 187 countries and cellular telephone data are downloaded from the ITU website (see pp. 228 - 229 of "Measuring the Information Society: 2013" by the International Telecommunication Union (ITU) at http://www.itu.int/en/ITU-D/Statistics/Pages/publications/mis2013.aspx) for a cross-section of 157 countries. When these two data sets are merged there are 156 observations. (The ITU data include Macao, China as an observation but the UN data do not. All of the other ITU observations are also in the UN data.) The figure to the right shows a fairly linear association for the entire cross-section of 156 countries. It also reports the R^2, SST, SSR, and SSE.



The graphs to the right use only the 34 member nations of the OECD. The last graph excludes three OECD members: Turkey and Mexico, which have rather low HDI's, and Canada with has the lowest mobile-cellular subscriptions per 100 inhabitants of all OECD members. Finally, the raw data for the OECD member nations are reproduced below. These data are observational, cross-sectional data.





country_name	country_code	hdi_2012	mobile_tel_subs_per_100_2012
Australia	AUS	0.938	106.2
Austria	AUT	0.895	161.2
Belgium	BEL	0.897	119.4
Canada	CAN	0.911	75.7
Chile	CHL	0.819	138.5
Czech Republic	CZE	0.873	122.8
Denmark	DNK	0.901	118
Estonia	EST	0.846	154.5
Finland	FIN	0.892	172.5
France	FRA	0.893	98.1
Germany	DEU	0.92	131.3
Greece	GRC	0.86	116.9
Hungary	HUN	0.831	116.4
Iceland	ISL	0.906	105.4
Ireland	IRL	0.916	107.1
Israel	ISR	0.9	119.9
Italy	ITA	0.881	159.5
Japan	JPN	0.912	109.4
Korea (Republic of)	KOR	0.909	110.4
Luxembourg	LUX	0.875	145.5
Mexico	MEX	0.775	86.8
Netherlands	NLD	0.921	117.5
New Zealand	NZL	0.919	110.3
Norway	NOR	0.955	115.5

Poland	POL	0.821	132.7
Portugal	PRT	0.816	115.1
Slovakia	SVK	0.84	111.2
Slovenia	SVN	0.892	110.1
Spain	ESP	0.885	108.3
Sweden	SWE	0.916	122.6
Switzerland	CHE	0.913	135.3
Turkey	TUR	0.722	90.8
United Kingdom	GBR	0.875	130.8
United States	USA	0.937	98.2

(a) Compare and contrast the R^2 across the three scatter diagrams. What explains any differences?

(b) Compare and contrast the SST, SSR, and SSE across the three graphs. What explains any differences?

(8) Each year the U.S. Department of Energy releases a fuel economy guide to inform consumers about the fuel economy and greenhouse gas emissions associated with vehicles (cars, vans, etc.) released that year. Further, on the website (www.fueleconomy.gov), they release detailed data for each make and model of vehicle each year. Consider the most recent data on 1,250 makes and models in 2015 (e.g. Ford Focus with automatic transmission, Honda Civic with manual transmission). These data include: the fuel economy (FE) in city driving in miles per gallon (MPG), the FE in highway driving in MPG, and a green house gas (GHG) emissions rating on a scale from 1 to 10 where 1 is the worst and 10 is the best.

(a) Which kind of data are these? How many observations? How many variables?

(b) Consider this tabulation of the GHG rating and the fact that the sample mean is 5.3 and the s.d. is 1.6. What is the approximate shape of the distribution? Suppose a vehicle had a GHG rating of 3: what would the *standardized* rating be? How would you interpret it?

GHG Rating	Freq.	Percent	Cum.
1	27	2.16	2.16
2	23	1.84	4.00
3	83	6.64	10.64
4	244	19.52	30.16
5	359	28.72	58.88
6	206	16.48	75.36
7	199	15.92	91.28
8	86	6.88	98.16
9	18	1.44	99.60
10	5	0.40	100.00
Total	1,250	100.00	

(c) Consider this variance-covariance matrix for city and highway fuel economy. Fully interpret all numbers (which will include computing other numbers using these).

	city_fe	hwy_fe
city_fe	30.4814	
hwy_fe	31.1006	38.2749

(d) Consider these OLS regression results. Fully interpret all numbers.

```
city_fe-hat = -2.53 + 0.81*hwy_fe; R-squared = 0.83; n = 1,250
```