

17.871, Political Science Lab
Spring 2004
Problem set # 3: Bivariate regression

Handed out: March 4, 2004
Returned: March 16, 2004

General note: when the question is marked with an asterisk (*) I want you to include a log file of the STATA output you used to produce the result. Also include any other calculations you made to get the answer.

- 1*. The following table reports the total level of federal spending from FY 1980 to FY 2003, in nominal (i.e., non-inflation-adjusted) dollars. Letting Y denote spending and X the year, run the following two regressions:

$$Y_t = \beta_0 + \beta_1 X_t + u_t$$

and

$$\ln Y_t = \beta_0 + \beta_1 X_t + u_t$$

Estimate each regression equation and report how you would interpret β_1 in each case.

Fiscal Year	Expenditures (\$millions)
1980	590,941
1981	678,241
1982	745,743
1983	808,364
1984	851,853
1985	946,396
1986	990,430
1987	1,004,082
1988	1,064,455
1989	1,143,646
1990	1,253,165
1991	1,324,369

1992	1,381,655
1993	1,409,489
1994	1,461,877
1995	1,515,802
1996	1,560,535
1997	1,601,250
1998	1,652,585
1999	1,701,891
2000	1,788,773
2001	1,863,770
2002	2,010,970
2003	2,157,637

2*. Use the data set /mit/course_number/Examples/residual_vote_1996.dta to answer the following questions. There is a description of the variables immediately below.

- 2.1 What was the correlation coefficient between the median house value in a county and the residual vote rate of that county? Does the correlation coefficient you calculated fairly describe this relationship? (Hint: Think about reasonable transformations of the variables.)
- 2.2 Run the regression associated with the relationship between median house value and residual vote rate. If you think the variables need to be transformed to better reflect the relationship, do that. Write a sentence that describes how you would interpret this regression.

Variable name	description
stabbr	state postal abbreviation
area_name	county or similar state subdivision
median_val	median value of single family houses (uits: \$100,000)
pdrop	fraction of ballots with no vote for president (i.e., residual vote)

3. In a recent piano competition, information was gathered from each of the contestants concerning the average number of hours each day each person practiced. Let us say you were a hovering stage parent who wanted to increase the future competition scores of your child. You run a regression with this data, with the competition score as the dependent variable and number of practice hours as the independent variable. Would the resulting slope coefficient from this regression tell you how much better your child would perform (on average, of course) if she or he were to practice another hour each day? Why or why not?

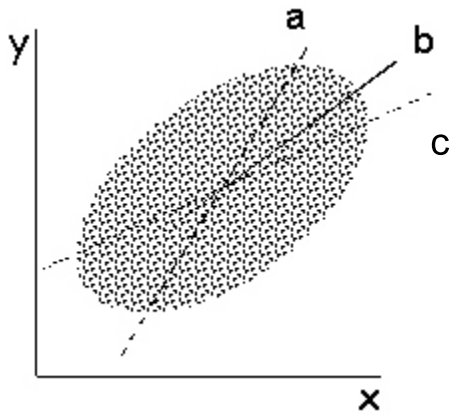
4. Below are the data from the online article we discussed in class, that related how worried voters were about the state of the economy (relative to how worried they were about international affairs) with how well the incumbent president did when he ran for reelection.
 - 4.1 Calculate by hand the appropriate regression that predicts how well presidents do in these circumstances. Show the components that go into calculating the regression coefficients

 - 4.2* Run this regression using STATA. Analyze the residuals and report on any elections that appear not to obey the regression.

 - 4.3 Using the previous regression, calculate the estimated incumbent win/lose margin for George Bush in 2004.

Year	Net Mention of the economy as the most important problem facing the country (economy minus international)	Incumbent win/lose margin in the general election	Incumbent
1980	51	-10	Carter
1976	48	-2	Ford
1992	34	-5	Bush I
1984	24	18	Reagan
1996	10	8	Clinton
2004	4	?	Bush II
1972	-10	23	Nixon
1964	-12	23	Johnson

5. Lester Maddox was an infamous governor of Georgia from 1967–1971, who was an avowed segregationist. However, the correlation between the (percentage of) votes Maddox received in each Georgia county with the percentage of the population that was Black was .54. What would account for a correlation like this? Does this suggest that Blacks were voting for Maddox more than Whites?
6. A foreign assistance program was designed as follows: Communities in an undeveloped country were ranked according to their levels of infant mortality. The ten communities with the highest infant mortality rates were chosen to receive health clinics; the ten communities with the lowest infant mortality rates were chosen as a control—they weren't given clinics. Five years later the ten communities that previously had had the highest infant mortality rates showed a decrease in mortality; the ten communities that had had the lowest mortality rates saw an increase. Program sponsors declare the clinics a success. Why is this an unwise inference?
7. Three lines are drawn across the following scatterplot. One is the standard deviation line, one is the line describing the regression of y on x , and the other is the line describing the regression of x on y . Which is which?



- 8*. You are interested in studying how much money industrialized countries spend each year on public health (i.e., on things like vaccinations, sanitation, etc.). You suspect that public health spending will be highly correlated with total health spending. Your strategy, in your exploratory analysis phase, is to examine the residuals from a regression of public health spending on total health spending, to identify the over- and under-providers of public health spending.

The following data, taken from the 2000 *Statistical Abstract of the United States*, shows the relevant information. Identify the three countries that deviate the most from the regression prediction on the positive side and the three that deviate the most on the negative side.

Country	Total health expenditures (pct. of GDP)	Public health expenditures (pct. Of GDP)
United States	12.9	5.8
Australia	8.6	6
Austria	8	5.8
Belgium	8.6	6.1
Canada	9.3	6.5
Czech Republic	7.1	6.5
Denmark	8.3	6.8
Finland	6.9	5.3
France	9.3	7.1
Germany	10.3	7.8
Greece	8.4	4.7
Hungary	6.8	5.6
Iceland	8.4	7
Ireland	6.8	5.2
Italy	7.7	5.5
Japan	7.5	5.8
Korea, South	5.1	2.4
Luxembourg	6	5.5
Mexico	5.3	2.6
Netherlands	8.7	6
New Zealand	8.1	6.3
Norway	8.6	7.1
Poland	6.4	4.2
Portugal	7.7	5.1
Spain	7	5.4
Sweden	7.9	6.6

Switzerland	10.4	7.6
Turkey	4.8	3.5
United Kingdom	6.8	5.7

9. If a linear regression of Y on X results in a slope coefficient of 2.0 and intercept of 1.0, what will the new slope and intercept coefficients be under the following circumstances:

- 9.1 Y is multiplied by 4.
- 9.2 X is multiplied by 4.
- 9.3 4 is added to Y
- 9.4 4 is added to X