Department of Economics UN3412
Columbia University Fall 2021

Problem Set 1
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw
For all sections

"Calculator" was once a job description. This problem set gives you an opportunity to do some calculations on the relation between smoking and lung cancer, using a (very) small sample of five countries. The purpose of this exercise is to illustrate the mechanics of ordinary least squares (OLS) regression. You will calculate the regression "by hand" using formulas from class and the textbook. For these calculations, you may relive history and use long multiplication, long division, and tables of square roots and logarithms; or you may use an electronic calculator or a spreadsheet.
The data are summarized in the following table. The variables are per capita cigarette consumption in 1930 (the independent variable, "*X*") and the death rate from lung cancer in 1950 (the dependent variable, "*Y*"). The cancer rates are shown for a later time period because it takes time for lung cancer to develop and be diagnosed.

Observation #	**Country**	**Cigarettes consumed per capita in 1930 (*X*)**	**Lung cancer deaths per million people in 1950 (*Y*)**
1	Switzerland	530	250
2	Finland	1115	350
3	Great Britain	1145	465
4	Canada	510	150
5	Denmark	380	165

Source: Edward R. Tufte, *Data Analysis for Politics and Management*, Table 3.3.

1. (40p) Use a calculator, a spreadsheet, or "by hand" methods to compute the following; refer to the textbook for the necessary formulas. (*Note*: please do not use functions like 'Correl' in Excel, we want you to calculate sample slope with the formula, if you use a spreadsheet, attach a printout)

(a) (5p) The sample means of X and Y, $\bar{X}$ and $\bar{Y}$.
(b) (6p) The standard deviations of X and Y, s_X and s_Y.
(c) (6p) The correlation coefficient, r, between X and Y.
(d) (6p) $\hat{\beta}_1$, the OLS estimated slope coefficient from the regression $Y_i = \beta_0 + \beta_1 X_i + u_i$
(e) (5p) $\hat{\beta}_0$, the OLS estimated intercept term from the same regression.
(f) (6p) $\hat{Y}_i$, $i = 1,\ldots, n$, the predicted values for each country from the regression
(g) (6p) $\hat{u}_i$, the OLS residual for each country.

2. (10p) On graph paper or using a spreadsheet, graph the scatterplot of the five data points and the regression line. Be sure to label the axes, the data points, the residuals, and the slope and intercept of the regression line.

3. (30p) Suppose that you were told that average hourly wages in State A and State B are different. However, you have your own suspicion that it could be the same. You suggest to your brother, who is attending a school in State B to collect data on hourly wage as part of your project in ECON3412. You do the same for State A as your school is located in State A. The accompanying table shows your and your brother's findings.

Average Hourly Wages in State A and State B, in dollar

State A			State B		
$\overline{Y}_{State\ A}$	$s_{State\ A}$	$n_{State\ A}$	$\overline{Y}_{State\ B}$	$s_{State\ B}$	$n_{State\ B}$
57.8	3.9	155	58.7	4.2	157

- **(a)** (6p) Let your null hypothesis be that there is no difference in the average hourly wage in the two states. Specify the alternative hypothesis.
- **(b)** (6p) Find the difference in average hourly wage and the standard error of the difference.
- **(c)** (6p) Generate a 95% confidence interval for the difference in average hourly wage.
- **(d)** (6p) Calculate the *t*-statistic for comparing the two means. Is the difference statistically significant at the 1% level? Which critical value did you use? Why would this number be smaller if you had assumed a one-sided alternative hypothesis? What is the intuition behind this?
- **(e)** (6p) What about at 5% level?

4. (20p) Suppose that the following table is the joint probability distribution of two random variables X and Y:

		X			
		-2	**0**	**2**	**3**
Y	**2**	**0.27**	**0.08**	**0.16**	**0.2**
	5	**0.1**	**0.04**	**0.1**	**0.05**

(a) (4p) Find the marginal PDF of X when x=-2, 0, 2, and 3.
(b) (4p) Find the marginal PDF of Y when y=2 and 5.
(c) (4p) Find the conditional PDF of x=-2 and 3 given that y=2 has occurred.
(d) (4p) Find the conditional PDF of y=2 and 5 given that x=3 has occurred.
(e) (4p) State conditional PDF in terms of Joint and marginal PDFs.

Following questions will not be graded, they are for you to practice and will be discussed at the recitation:

1. [Practice question, not graded] SW 2.3

	Rain (X=0)	No Rain (X=1)	Total
Long Commute (Y=0)	0.15	0.07	0.22
Short Commute (Y=1)	0.15	0.63	0.78
Total	0.30	.70	1.00

Using the random variables X and Y from Table 2.2 (given above), consider two new random variables W = 3 + 6X and V = 20 – 7Y. Compute:

a) E(W) and E(V).

b) $\sigma^2{}_W$ and $\sigma^2{}_{V.}$

c) $\sigma_{W,V}$ and Corr(W,V).

2. [Practice question, not graded] SW 2.6

The following table gives the joint probability distribution between employment status and college graduation among those either employed or looking for work (unemployed) in the working age US population, based on the 1990 US Census.

	Unemployed (Y=0)	Employed (Y=1)	Total
Non-college grads (X=0)	0.045	0.709	0.754
College grads (X=1)	0.005	0.241	0.246
Total	0.050	0.950	1.000

a. Compute E(Y).

b. The unemployment rate is the fraction of the labor force that is unemployed. Show that the unemployment rate is given by 1-E(Y).

c. Calculate the E(Y|X=1) and E(Y|X=0).

d. Calculate the unemployment rate for (i) college graduates and (ii) non-college graduates.

e. A randomly selected member of this population reports being unemployed. What is the probability that this worker is a college graduate? A non-college graduate?

f. Are educational achievement and employment status independent? Explain.

3. [Practice question, not graded] SW 2.14 [Hint: Use SW Appendix Table 1.]

In a population E[Y] = 100 and Var(Y) = 43. Use the central limit theorem to answer the following questions:

a. In a random sample of size n = 100, find Pr($\bar{Y}$ ≤101)

b. In a random sample of size n = 165, find Pr($\bar{Y}$ >98)

c. In a random sample of size n = 64, find Pr(101 ≤ $\bar{Y}$ ≤103)

4. [Practice question, not graded] SW 3.12

To investigate possible gender discrimination in a firm, a sample of 100 men and 64 women with similar job descriptions are selected at random. A summary of the resulting monthly salaries are:

	Avg. Salary ($\bar{Y}$)	Stand Dev (of Y)	n
Men	\$3100	\$200	100
Women	\$2900	\$320	64

a. What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that wages of men and women are different? (To answer this question, first state the null and alternative hypothesis; second, compute the relevant t-statistic; and finally, use the p-value to answer the equation.)

b. Do these data suggest that the firm is guilty of gender discrimination in its compensation politics? Explain.

5. [Practice question, not graded] SW 2.10 [Hint: Use SW Appendix Table 1.]

Compute the following probabilities:

a. If Y is distributed N(1,4), find Pr(Y≤3).

b. If Y is distributed N(3,9), find Pr(Y>0).

c. If Y is distributed N(50,25), find Pr(40≤Y≤52).

d. If Y is distributed N(5,2), find Pr(6≤Y≤8)

6. [Practice question, not graded] SW 3.3

In a survey of 400 likely voters, 215 responded that they would vote for the incumbent and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely voters that preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents that preferred the incumbent.

a. Use the survey results to estimate p.

b. Use the estimator of the variance of $\hat{p}$, $\hat{p}$ (1 - $\hat{p}$)/n to calculate the standard error of your estimator.

c. What is the p-value for the test H0: p=0.5 vs. H1:p≠0.5?

d. What is the p-value for the test H0: p=0.5 vs. H1:p>0.5?

e. Why do the results from (c) and (d) differ?

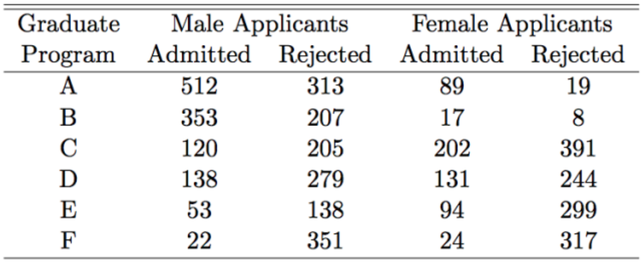
f. Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

7. Consider two events A and B with Pr(A) = 0.5 and Pr(B) = 0.9. Determine the maximum and minimum values of Pr(A ∪ B).

8. Assume that events A and B^c are independent. That is, Pr(A ∩ B^c) = Pr(A)Pr(B^c). Are events A and B also independent?

9. Let X and Y denote two random variables.
(i) Show that if at least one of X or Y has expectation equal to zero, then cov(X, Y) = E[XY].
(ii) Show that cov(αX + a, βY + b) = αβcov(X, Y) for constants (α, β, a, b).
(iii) Show that var(X + Y) = var(X) + var(Y) + 2cov(X, Y).
(iv) Show that the population correlation coefficient is always between [−1, 1].

10. The following admission data are for the graduate program in the six largest majors at the University of California at Berkeley for the fall 1973 quarter.



Graduate	Male Applicants		Female Applicants	
Program	Admitted	Rejected	Admitted	Rejected
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317

(a) What is the overall probability of being admitted for males? For females? What is the standard deviation for males and for females?
(b) How would you write down the null and alternative hypotheses in order to test that the overall probability of admission is higher for men than for women?
(c) Conduct a t-test of the hypothesis from part (b) and report the p-value.
(d) Is the result significant at the 5% level? Does it provide evidence of discrimination?
(e) Committee chairpersons claim they are more likely to admit women than men. Is this claim true? Compute acceptance rates for men and women by graduate program.

(f) Do these data suggest that the university is guilty of gender discrimination in its admission policy? Explain briefly.