Correlation and regression.

1. ***Doll's ecological study of smoking and lung cancer.*** In 1955, Richard Doll published an ecological study of smoking and lung cancer. Smoking was measured as per capita cigarette consumption in 1930 (CIG). Lung cancer mortality per 100,000 person-years in 1950 (LUNGCA). Data are shown in the table below.

(A) Construct a scatterplot of the relation between cigarette consumption and lung cancer. Consider the form, direction, and strength of the relationship. Are there any outliers?

(B) Calculate the correlation coefficient for the problem. Interpret this statistic.
(C) Test the correlation coefficient for significance. Show all hypothesis testing steps (null hypothesis statement, test statistic, *P* value, conclusion).
(D) What % of LUNGCA is "explained" by CIG?

i	COUNTRY	CIG	LUNGCA
1	USA	1300	20
2	Great Brit	1100	46
3	Finland	1100	35
4	Switzerland	510	25
5	Canada	500	15
6	Holland	490	24
7	Australia	480	18
8	Denmark	380	17
9	Sweden	300	11
10	Norway	250	9
11	Iceland	230	6

2. ***Sodium and blood pressure***. Data (n = 10) on daily SODIUM intake (mg) and systolic blood pressure (BP; mm Hg) are shown below.

(A) Which variable is the explanatory variable in this analysis? Which is the response variable?
(B) Construct a scatter plot of these data. *Discuss* your plot by considering its form, direction of association, and strength of association. Are there any outliers?
(C) Compute *r*. Interpret this statistic.
(D) What % of BP is explained by SODIUM?
(E) Test the correlation for significance. Show all hypothesis testing steps (null hypothesis, test statistic, *P* value, conclusion).

i	SODIUM	BP
1	6.8	154
2	7.0	167
3	6.9	162
4	7.2	175
5	7.3	190
6	7.0	158
7	7.0	166
8	7.5	195
9	7.3	189
10	7.1	186

3. ***Gravid iguanas.*** Data on post-partum body weight (kilograms) and the number of eggs produced by gravid iguanas are shown below.

(A) Construct a scatter plot of the data. (Make certain you put the explanatory variable on the horizontal axis.) Interpret your plot.
(B) Calculate the correlation coefficient. Interpret this statistic.
(C) Test the correlation coefficient for statistical significance.

i	WEIGHT	EGGS
1	0.90	33
2	1.55	50
3	1.30	46
4	1.00	33
5	1.55	53
6	1.80	57
7	1.50	44
8	1.05	31
9	1.70	60

4. ***Graduation rates at Big Ten universities***. The most reliable factor that predicts graduate is scholastic aptitude and motivation. To explore quantify this fact, a researcher collects data on many factors. Data is given below. Graduation rates by university (percentage of students graduating within 5 years of entry) are stored in the variable UPERCENT. The average ACT scores of incoming freshman at is the predictor variable for this analysis.

(A) Plot these data. Interpret your plot.
(B) Calculate *r*. Interpret this statistitic.
(C) Test it for statistical significance. Interpret your results.
(D) Calculate r^2. What does this tell you about the variability of graduation rates?

UPERCENT	ACT
76.2	27
57.6	24
55.4	24
59.7	23
86.0	28
46.2	22
66.7	23

5. ***Maternal mortality and health care during birth***. This study explored the relation between the percentage of births attended by physicians, nurses, and midwives (ATTENDED) and maternal mortality per 100,000 live births (MAT_MORT). The values for a random sample of 11 countries are shown below Data are a sample from Pagano & Gauvreau (2000, p. 407) as originally published in United Nation's Children's Fund (1994).

COUNTRY	ATTENDED	MAT_MORT_
Bangladesh	5	600
Chile	98	67
Iran	70	120
Kenya	50	170
Nepal	6	830
Netherlands	100	10
Nigeria	37	800
Pakistan	35	500

Panama	96	60
United States	99	8
Vietnam	95	120

(A) What is the independent variable in this analysis? What is the dependent variable?
(B) Plot the data as a scatterplot. Interpret what you see (form, direction, strength, outliers if any).
(C) Calculate *r*. Interpret this statistic.
(D) Test the correlation for statistical significance. Show all hypothesis testing steps.
(E) Identify lurking variables that may confound and observed relationship. Explain how confounding may occur.

6. ***Need and demand for mental health care.*** This example uses data from a 1854 study on mental health care in the fourteen counties in Massachusetts in the prior century. The study conducted by Edward Jarvis. Jarvis, then president of the American Statistical Association. The explanatory variable is the reciprocal of the distance (in miles^{-1}) to the nearest mental healthcare center (REC_DIST). The response variable is the percent of patients cared for in the home (PHOME). The relation between the percentage of patients cared for at home and distance to the nearest health care center remains important today--it is still recommended that numerous small mental hospitals be erected at scattered locations rather than having one large central facility

(A) Create a scatterplot of the relation between PHOME and REC_DIST. Describe the relationship. Are there any outliers?
(B) Calculate the correlation coefficient using all 14 data points.

COUNTY	PHOME	REC_DIST
BERKSHIRE	77.00	.01031
FRANKLIN	81.00	.01613
HAMPSHIRE	75.00	.01852
HAMPDEN	69.00	.01923
WORCESTER	64.00	.05000
MIDDLESEX	47.00	.07143
ESSEX	47.00	.10000
SUFFOLK	6.00	.25000
NORFOLK	49.00	.07143
BRISTOL	60.00	.07143
PLYMOUTH	68.00	.06250
BARNSTABLE	76.00	.02273
NANTUCKET	25.00	.01299
DUKES	79.00	.01923

7. Atherosclerotic heart disease as a function of fat calories. Following World War II, it became clear that northern European countries with high dietary fat consumption were experiencing notable increases in what was then called degenerative heart disease. Data in this exercise are a fictionalized version of data from early ecological studies reported by Keys (1952, also see EKS p. 195). Data for calories from fat as a % of total calories (FAT_CAL) and CHD mortality per 1000 50- to 59-year-olds are:

COUNTY	FAT_CAL	CHD
Japan	8	0.5
Italy	20	1.4
England	33	3.8
Australia	36	5.5

Canada	37	5.7
USA	39	7.1

(A) Which of the variables in this data set is the independent variable? Which is the dependent (response) variable?
(B) Plot the data.
(C) Can the relation be described with a straight line?
(D) Calculate the correlation coefficient using the data.