MATHS. Biochemistry degree

SHEET 2 : LINEAR REGRESSION

1. In a lab experiment, we are given five samples of glycogen, to which we apply a known concentration of the enzyme glucokinase (in mmol/l) and check the reaction rate (in μmol/min):

concentr glucokinase	1	2	3	0′25	0′5
reaction rate	18	35	60	8	10

   (a) Decide whether the reaction rate depends linearly on the concentration of glucokinase.

   (b) Give a prediction for the reaction rate in a sample with 2'5 mmol/l of glucokinase.

2. The Lowry method is used in biochemistry to determine the protein concentration in a solution, by adding a certain reactive which produces a color change. More precisely, the blue intensity Y depends linearly on the protein concentration X. We wish to calibrate a lab *colorimeter*, so we measure the blue intensities of 8 known protein concentrations (in μg/ml), obtaining the data

Concentr	0	25	50	100	150	200	250	300
Intensity	0	0.070	0.146	0.240	0.359	0.461	0.562	0.683

   (a) For these data, find the regression line $Y = a + bX$. Find the covariance and the coefficient r.

   (b) Suppose that an unknown sample exhibits a blue intensity level $y = 0.160$. Give an estimation of the protein concentration for this sample.

3. The number of manatees found dead in the Florida shores increases every year. This maybe related with the presence of motorboats in their waters, which hit or kill many of them. The next table shows, for several years, the number of motorboat licenses (in thousands) vs the number of dead manatees.

Year	Licenses	Manatees	Year	Licenses	Manatees
1997	447	13	2004	559	34
1998	460	21	2005	585	33
1999	481	24	2006	614	33
2000	498	16	2007	645	39
2001	513	24	2008	675	43
2002	512	20	2009	711	50
2003	526	15	2010	719	47

   a) Draw a scatter plot for the number of dead manatees vs the number of motorboat licenses. Explain what you see about the relation of these two variables (you must decide which variable is dependent and which independent).

   b) Find the regression line for the number of dead manatees vs the number of motorboat licenses. Evaluate if there is a strong linear relation between these two variables.

   c) Florida authorities want to ensure that the number of dead manatees does not go beyond 40 individuals. Give a prediction for the number of yearly motorboat licenses that the Florida state should allow.

4. The following three data sets were obtained by the statistician Frank Anscombe to illustrate the necessity of graphically representing data before doing any computations. Observe that the three data sets have the *same* regression line and coefficient r.

   Data set A:

X	10	8	13	9	11	14	6	4	12	7	5
Y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68

   Data set B:

X	10	8	13	9	11	14	6	4	12	7	5
Y	9,14	8,14	8,74	8,77	9,26	8,10	6,13	3,10	9,13	7,26	4,74

   Data set C:

X	8	8	8	8	8	8	8	8	8	8	19
Y	6,58	5,76	7,71	8,84	8,47	7,04	5,25	5,56	7,91	6,89	12,50

a) Find the regression line and correlation coefficient r for each of the data sets, and show that they are approximately equal.

b) For each of data sets, draw the scattering plots together with the corresponding regression lines.

c) In which data sets would you use the regression line to predict a value for Y, given that $X = 14$?

5. Suppose that two variables x and y have respective means 7 and 10, and standard deviations 3 and 2. Suppose also that the correlation is $0'7$. Find the equation of the regression lines for (x, y) and for (y, x), and compute the LSE in each case.

6. The lifespan of a certain species of insect depends on the humidity level of their living habitat. To study this dependence we take samples from 10 different habitats, and measure the variables X="Average humidity" and Y="Lifespan". We obtain, respectively, the mean values 59% humidity and 28'7 days. The remaining data are given by

$$\sum x_i^2 = 35140; \qquad \sum y_i^2 = 8573; \qquad \sum x_i y_i = 17260$$

a) Find the regression line for Y in terms of X, and evaluate the goodness of fit.

b) Predict the lifespan for a habitat with humidity level 65%.

7. In populations with limited resources, the *logistic model* predicts that the annual growth rate per capita G depends linearly on the number of individuals N, according to the rule

$$G = G_0\Big(1 - \frac{N}{N_\infty}\Big),$$

for suitable constants G_0 = (initial growth rate) and N_∞ (maximum capacity of the population). We wish to apply this model to the growth of the world population, for which we have the data:

year	1965	1970	1975	1980	1985	1990	1995	2000	2005
N ($\times 10^9$)	3,335	3,692	4,068	4,435	4,831	5,264	5,674	6,071	6,454
G (en %)	2,648	2,421	2,308	2,067	2,051	2,055	1,810	1,632	1,484

(i) Find the regression line for $G = a + bN$.

(ii) Represent graphically the data and their regression line, and evaluate the goodness of fit.

(iii) From the values of a, b obtained in (i), give an estimate of the constats G_0 and N_∞.

8. In biochemical reactions catalyzed by an enzyme, the reaction rate V depends on the concentration of substrate $[S]$, according to the Michaelis-Menten equation

$$V = \frac{v_{\max}\,[S]}{k + [S]}, \tag{1}$$

where $k, v_{\max}$ are positive parameters. In lab experiments one wants to estimate these parameters. To do so, one possibility is to find the regression line of $1/[S]$ and $1/V$, since (1) can also be written as

$$\frac{1}{V} = \frac{k}{v_{\max}}\,\frac{1}{[S]} + \frac{1}{v_{\max}}.$$

Suppose that, for a certain reaction in our lab we obtain the data

$[S]$	0	1	5	15	40	65	80	90	100
V	0	37	75	91	96	98	98.5	99	100

(i) Write a table with the values of $1/[S]$ and $1/V$.

(ii) Find the regression line $\frac{1}{V} = a + b\frac{1}{[S]}$, and evaluate the goodness of fit.

(iii) Use the previous steps to estimate the parameters k and $v_{\max}$.

(iv) Predict the reaction rate V when $[S] = 50$