ARTIFICIAL INTELLIGENCE

Lab8. *Correlation and regression*

Caution! All the data are available in respective *csv* files!

1. The following sample observations have been obtained by a chemical engineer investigating the relationship between weight of final product Y (in pounds) and volume of raw materials X (in gallons):

X (volume of raw materials)	14	23	9	17	10	22	5	12	6	16
Y (weight of final product)	68	105	40	79	81	95	31	72	45	93

   (a) compute the sample covariance between the weight of final product and the volume of raw materials — 138.49

   (b) compute and interpret the sample correlation coefficient — 0.895

   (c) draw the data on a scatter plot

   (d) determine regression line of the final yield of a product with respect to the volume of raw materials used for production; represent regression line on the graph with scatter plot — $y = 22.4 + 3.62x$

   (e) describe, how the final yield of a product will change if the volume of raw materials increases by 1 gallon — increase by 3.62 pounds

   (f) compute, what will be the final yield of a product if the volume of raw materials is equal to 15 gallons; — 77

   (g) compute, what will be the final yield of a product if the volume of raw materials is equal to 20 gallons — 95

   (h) compute, how good is approximation by linear function; — 80%

   (i) test the hypothesis about significance of regression coefficient. Assume 5% of significance level. What is the interpretation of the decision? — $F = 32.22$, reject H_0

2. Arsenic is found in many ground-waters and some surface waters. Recent health effects research has prompted the Environmental Protection Agency to reduce allowable arsenic levels in drinking water so that many water systems are no longer compliant with standards. This has spurred interest in the development of methods to remove arsenic. The accompanying data on pH (*X*) and percentage arsenic removed (*Y*) by a particular process was collected:

X	7.01	7.11	7.12	7.24	7.94	7.94	8.04	8.05	8.07	8.90	8.94	8.95	8.97	8.98	9.85	9.86	9.86	9.87
Y	60	67	66	52	50	45	52	48	40	23	20	40	31	26	9	22	13	7

   (a) compute the covariance between the pH of water and percentage arsenic removed — -18.32

   (b) compute and interpret the correlation coefficient — -0.95

   (c) draw the data on a scatter plot

   (d) determine regression line of the percentage arsenic removed with respect to the pH of water; present the line on the graph with scatter plot — $y = 190.27 - 18.03x$

   (e) describe, how the percentage of arsenic removed will change if the pH of water increases by 1 — decreases by 18.03 (%)

   (f) compute, what will be the percentage of arsenic removed if the pH of water is equal to 7.5 — 55

   (g) compute, what will be the percentage of arsenic removed if the pH of water is equal to 9 — 28

   (h) compute, how good is approximation by linear function; — 90%

   (i) test the hypothesis about significance of regression coefficient. Assume 1% of significance level. What is the interpretation of the decision? — $F = 149.7$, reject H_0

3. Variations in clay brick masonry weight have implications not only for structural and acoustical design but also for design of heating, ventilating, and air conditioning systems. The following table represents X – mortar air content (%) and Y – mortar dry density (lb/ft^3):

X	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0	15.3	16.2	17.8	18.7	19.7	20.6	25.0
Y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1	111.3	107.2	108.9	107.8	111.0	106.2	105.0

(a) compute the covariance between the mortar air content and mortar dry density -26.6

(b) compute and interpret the correlation coefficient -0.87

(c) draw the data on a scatter plot

(d) determine regression line of the mortar dry density with respect to the mortar air content; present the line on the same graph as scatter plot $y = 126.25 - 0.92x$

(e) describe, how the mortar dry density will change if the mortar air content increases by 1%

decreases by 0.92 (lb/ft^3)

(f) compute, what will be the mortar dry density if the mortar air content is equal to 11% 116

(g) compute, what will be mortar dry density if the mortar air content is equal to 23%; 105

(h) compute, how good is approximation by linear function; 75%

(i) test the hypothesis about significance of regression coefficient. Assume 1% of significance level. What is the interpretation of the decision? $F = 39.51$, reject H_0

4. It was decided to test the processor temperature [^{o}C] depending on the computer operation time [in h]. The results are presented in the table:

X (operation time)	1	4	7	8	10	13	18	22	25	28
Y (temperature)	35	45	59	77	62	50	39	47	54	71

a) Determine and interpret covariance and correlation coefficient between temperature and the operation time.
b) On the basis of the scatter plot, verify possible relationship between temperature and operation time.
c) Determine polynomial regression for the temperature depending on operation time; try various degrees of the polynomial.
d) Add a polynomial regression functions to an existing plot.
e) Verify significance of regression; assume significance level 5%.
f) Predict the temperature after 15 and 48 hours.
g) Determine and interpret the coefficient of determination.

Caution! To determine square polynomial regression use function `lm(Y~X+I(X^2))`