SOLUTIONS TO HOMEWORK: REGRESSION

SOLVE THE FOLLOWING PROBLEMS IN TWO WAYS:
(a) USING CALCULATOR
(b) USING MS EXCEL

Write out the regression equation. What is the correlation coefficient? What is r-square (coefficient of determination).

In simple regression (one X variable and one Y variable), there are three ways to test for significance. All three tests are equivalent and give the same results. You can test the regression for significance by examining the F-value in the computer printout. Or, you can test the correlation coefficient for significance using a t-test. Or, you can test the slope term (b_1) for significance using a t-test (this is also on the computer printout). You should always do one test to make sure that you are not looking at a chance relationship that is meaningless. Even when a relationship is statistically significant, it may be of little practical importance if the correlation coefficient is low, say, below .30.

Problem 1: A researcher is interested in determining whether there is a relationship between number of packs of cigarettes smoked per day and longevity (in years). n=10.

# packs of cigarettes smoked (X)	Longevity (Y)
0	80
0	70
1	72
1	70
2	68
2	65
3	69
3	60
4	58
4	55

$\sum X = 20$; $\sum Y = 667$; $\sum XY = 1247$; $\sum X^2 = 60$; $\sum Y^2 = 44{,}983$

Answer: Longevity = 75.4 – 4.35 (Packs of cigarettes smoked)
The regression is statistically significant (F-value = 26.18, p =.0009); r = -.875 (negative relationship); r^2 = 76.6%; Every pack a person smokes per day results on average in a loss of 4.35 years of life. Someone who does not smoke, is expected to live for 75.4 years.

Problem 2: A researcher is interested in determining whether there is a relationship between advertising and sales for her firm. n = 11 sales areas.

advertising in $thousands(X)	Sales in millions(Y)

1	0
1	1
2	4
4	3
5	5
6	4
6	7
6	8
7	9
10	9
10	7

$\sum X = 58$; $\sum Y = 57$; $\sum XY = 383$; $\sum X^2 = 404$; $\sum Y^2 = 391$

Answer: Sales in millions = .75 + .84 (Advertising in thousands of dollars)
The F-value is 23.62 and p=.000896. Thus, the correlation and regression are significant; r = +.851; r^2 = 72.4%; Every $1000 of advertising results in increased sales of .84 million ($840,000). If the company does 0 advertising, they can expect sales of .75 million ($750,000).

Problem 3: A researcher is interested in determining whether there is a relationship between price and quantity demanded for her firm. n = 11.

Price(X)	Q-demanded(Y)
2	95
3	90
4	84
5	80
6	74
7	69
8	62
9	60
10	63
11	50
12	44

$\sum X = 77$; $\sum Y = 771$; $\sum XY = 4{,}864$; $\sum X^2 = 649$; $\sum Y^2 = 56{,}667$

Answer: Quantity demanded = 103.82 – 4.82 (Price)
The F-value is 313.66 and it is significant; p =.0000000265. Thus, the regression (and correlation) are significant; r = -.986; r^2 = 97.2%; Every increase of $1 in price results in decreased sales of 4.82 units. If the company charges $0 (i.e., free), they can expect sales of 103.82. Of course, this will not happen and illustrates the dangers of predicting

outside the range of the X values. Note that if the sign of b_1, the slope term is negative, r must negative.

Problem 4: A researcher is interested in determining whether there is a relationship between shelf space and number of books sold for her bookstore. n = 11.

Shelf Space in feet(X)	Books Sold(Y)
7.0	280
3.5	140
4.0	170
4.2	200
4.8	215
3.9	190
4.9	240
7.5	295
3.0	125
5.9	265
5.0	200

$\sum X = 53.7$; $\sum Y = 2{,}320$; $\sum XY = 12{,}070$; $\sum X^2 = 282.21$; $\sum Y^2 = 519{,}700$

Answer: Books sold = 29.77+ 37.10 (Shelf Space)
The F-value is 89.44 and it is significant; p =.00000568. Thus, the regression (and correlation) are significant; r = +.953; r^2 = 90.9%; Every increase of 1 foot in shelf space results in increased sales of 37.1 books. If the company provided no shelf space for a book, it can expect sales of 29.77. Of course, this may not happen and illustrates the dangers of predicting outside the range of the X values.

Problem 5: A researcher is interested in determining whether there is a relationship between grades and hours studied for statistics.

Hours studied(X)	Grade on final(Y)
1	20
2	30
4	40
7	60
6	65
7	70

8	80
9	90
8	95
10	100

$\sum X = 62$; $\sum Y = 650$; $\sum XY = 4{,}750$; $\sum X^2 = 464$; $\sum Y^2 = 49{,}150$

Answer: Grade = 8.92 + 9.045 (Hours Studied)
The F-value is 134.47 and it is significant; p =.00000278. Thus, the regression (and correlation) are significant; r = +.97; r^2 = 94.4%; Every additional hour studied results in a 9.045 increase in the grade. If a student does no studying, s/he can expect a grade of 8.92.

Problem 6: A researcher is interested in determining whether there is a relationship between number of police officer in a district and number of crimes

Number of Police Officers (X)	Number of Crimes (Y)
4	49
6	42
8	38
9	31
10	24
12	24
12	28
13	23
15	21
20	19
26	12
28	14

$\sum X = 163$; $\sum Y = 325$; $\sum XY = 3{,}593$; $\sum X^2 = 2{,}839$; $\sum Y^2 = 10{,}177$

Answer: Number of Crimes = 44.94 – 1.315 (number of Police Officers)
The F-value is 36.64 and it is significant; p =.000123. Thus, the regression (and correlation) are significant; r = -.886; r^2 = 78.6%; Every additional police results in a decrease in crime of 1.315. If there would be 0 cops in a district, they can expect 44.94 crimes. Of course, this will not happen and illustrates the dangers of predicting outside

the range of the X values. Note that if the sign of b_1, the slope term is negative, r must negative.

Problem 7: A researcher is interested in determining whether there is a relationship between education (in years) and net income (in thousands of dollars) studied for statistics.

Education in Years (X)	Income (in thousands) (Y)
9	20
10	22
11	24
11	23
12	30
14	35
14	30
16	29
17	50
19	45
20	43
20	70

$\sum X = 173$; $\sum Y = 421$; $\sum XY = 6{,}616$; $\sum X^2 = 2{,}665$; $\sum Y^2 = 17{,}129$

Answer: Income in Thousands = -11.02 + 3.20 (Education in years)
The F-value is 28.61 and it is significant; p =.000324. Thus, the regression (and correlation) are significant; r = +.86; r^2 = 74.1%; Every additional year of education studied results in a 3.20 thousand ($3,020)increase in income. An individual with 0 years of education, can expect an income of – 11.02 Thousand (-$11,020). This negative income probably indicates living off relatives or government.

Problem 8: A researcher is interested in determining whether there is a relationship between high school average and job performance ((the higher the number, the better the performance) at a certain company.

High school average (X)	Job performance (Y)
60	2
78	5
98	10
66	3
87	8
77	5

61	4
90	6
91	7
79	6
88	7
99	9
88	4
85	8
81	9

$\sum X = 1{,}228$; $\sum Y = 93$; $\sum XY = 7{,}932$; $\sum X^2 = 102{,}560$; $\sum Y^2 = 655$

Answer: Job Performance = -6.65 + .157 (High School Average) The F-value is 22.88 and it is significant; p =.000357. Thus, the regression (and correlation) are significant; r = +.799; r^2 = 63.8%; Every additional point of high school average, results in a .157 increase in job performance (10 points of high school average result in an increase of 1.57 in job performance rating). An individual with 0 high school average, can expect a negative rating of – 6.65. I have no idea what a negative rating means but this is not someone you would want to hire.

(Problem 9)

A researcher wants to see whether there is a relationship between the number of colds people get in one year and the average amount of vitamin C they consume.

Milligrams of Vitamin C (X)	Number of Colds (Y)
830	0
900	0
900	1
170	1
230	1
50	2
420	2
280	2
200	3
200	4
80	5
50	9

$\sum X = 4{,}310$; $\sum Y = 30$; $\sum XY = 5{,}050$; $\sum X^2 = 2{,}736{,}900$; $\sum Y^2 = 146$

Answer: Number Colds = 4.23 - .0048(milligrams of vitamin c). The F-value is 6.347 and it is significant; p =.030. Thus, the regression (and correlation) are significant; r = -.623; r^2 = 38.8%; Every milligram of Vitamin C, results in a .0048 decrease in number of colds (every 100 milligrams reduce the number of colds by .48). An individual with 0 milligrams of C, can expect 4.23 colds.

(Problem 10)

A researcher wants to see whether there is a relationship between the number of hours people exercise weekly and how long they live.

Hours Exercised (X)	Longevity (Y)
0	70
0	68
0	75
2	66
2	76
3	72
4	69
4	73
6	72
6	74
8	72
8	77
10	73
10	77
12	76
17	78
20	81
24	82
30	86
32	89

$\sum X = 198$; $\sum Y = 1506$; $\sum XY = 15{,}890$; $\sum X^2 = 3{,}782$; $\sum Y^2 = 114{,}048$

Answer: Longevity = 69.97 + .538 (Hours of Exercise) The F-value is 80.25 and it is significant; p =.000000047. Thus, the regression (and correlation) are significant;

r = +.904; r^2 = 81.7%; Every additional hour of exercise, results in a .538 year increase in longevity. An individual who does 0 hours of exercise weekly, can expect to live for 69.97 years.

ADDITIONAL TOPIC: TIME SERIES

This topic may not be covered. It is useful for those students interested in learning about trend lines.

Problem 10: A researcher is interested in drawing a trend line for the cost of a vital part using regression. Note: 1994 = 0; 1995 = 1; 1996 = 2; etc.

Time (X)	Cost of a part (Y)
0 (1994)	10
1	12
2	15
3	18
4	18
5	16
6 (2000)	19
7	22
8	25
9	30
10	35
11	32
12	31
13	35
14 (2008)	40

$\sum X = 105$; $\sum Y = 358$; $\sum XY = 3{,}075$; $\sum X^2 = 1{,}015$; $\sum Y^2 = 9{,}778$

Answer: Cost of Part = 9.64 + 2.03 (Time). The F-value is 194.098 and it is significant; p =.0000000034. Thus, the regression is significant;
r = +.968; r^2 = 93.7%; The cost in the base year (X=0) was $10; the cost of the part increases by $2.03 every year.