LINEAR REGRESSION

1. Choose correct answer:

1.1 Which of the terms belongs least to the other?

a) percentile
b) correlation
c) regression
d) prediction

1.2 About linear relationship between two variables…

a) … tells correlation more than regression equation
b) … tells correlation less than regression equation
c) … correlation and regression equation tell the same
d) … correlation and regression equation tell different information

2. The following questions ask about general aspects of linear regression:

2.1 What are the two important uses of regression analysis?

2.2 Why is the assumption of homoscedascity important?

2.3 What does it mean that regression line is based on least squares method?

2.4 Why $\sum(Y - Y')$ cannot be used as indicator of estimation error?

2.5 Describe how is correlation related to prediction/estimation error.

2.6 Does it apply that with increasing r the residual standard deviation increases as well?

2.7 Why researches have to be cautious if they use regression equation for estimation scores of a subject who has different characteristics than subjects included in the regression model?

3. Determine correct answers:

3.1 If $r = 0$, what is b?

3.2 If $r =1$, does it mean that in all pairs of correlated variables are both the values the same?

3.3 If $r =0.5$ and $z_X=2.0$, what is the estimate of the value Y, that is z_Y' ?

3.4 What is the percentile equivalent to the value z_Y' from previous question?

3.5 If $r =-0.6$ a $z_X=-1.5$, what is the estimate of value Y, that is z_Y' ?

3.6 If we predict Y from X and $s_X=s_Y=15$, does it apply that correlation (r) equals to regression coefficient b ?

3.7 If $s_Y = 10$ and $r = 0.6$, what is the residual standard deviation (s_e)?

3.8 If $r_{XY} = 0.5$, and performance of a subject in variable X corresponds with P_2, estimate the subject’s percentile in variable Y:

a) P_{50}
b) P_{75}
c) P_{16}
d) P_2

3.9 If s_Y = 10 and r = 0.6, what is the variance of predicted values Y' ?

3.10 If s_Y = 10 and r = 0.6, what is the residual standard deviation (s_e, resp. s_{res})?

4. If we assume bivariate normal distribution (that is, the joint distribution of two variables is normal):

4.1 If residual standard deviation s_e = 8, what percentage of real values of the dependent variable will differ from the estimated values by less than 8 points?

4.2 And what percentage of real values of the dependent variable will differ from the estimated values by more than 8 points?

4.3 What is the percentage of estimates which undervalue the real score by more than 8 points? Is it the same percentage as in the previous question?

5. Correlation between IQ scores of a parent and his child is approximately 0.5. Further we know that IQ mean is the same for children and parents (M=100) and SD as well (SD=15).

5.1 Estimate average IQ of children whose mothers have IQ = 130.

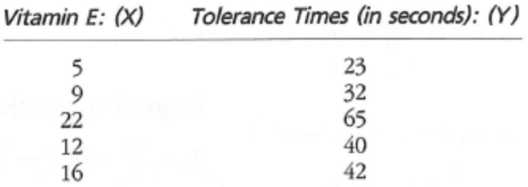
5.2 Estimate average IQ of children whose fathers have IQ = 90.

5.3 Estimate average IQ of children whose mothers have IQ = 100.

5.4 Average IQ of both parents correlates with children's IQ approximately 0.6. What will be the residual standard deviation (s_e, s_{res}) for prediction the children's IQ?

5.5 If s_{res} = 12, in what percentage of cases will the real IQ values differ from the predicted IQ values by more 12 points?

6. A researcher of pain is interested in prediction of time for which are subjects able to hold their hand in very cold water. He knows from preivous research that vitamin E consumed during last 12 hours correlates with tolerance of painful stimuli. The following table shows pairs of score for his sample:



Vitamin E: (X)	*Tolerance Times (in seconds): (Y)*
5	23
9	32
22	65
12	40
16	42

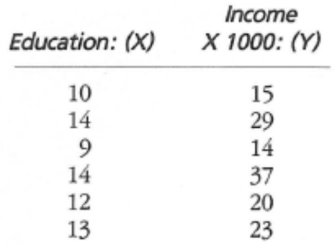
6.1 What is the slope b ?

6.2 What is the intercept a ?

6.3 What is the residual standard deviation s_e ?

6.4 What time would you predict for a person who took in the morning 16 units of vitamin E?

7. A sociologist is interested in prediction of year salary (Y) based on previous level of education (X). Level of education is defined as number of years in education. The following data were obtained from 6 subjects:



Education: (X)	*Income X 1000: (Y)*
10	15
14	29
9	14
14	37
12	20
13	23

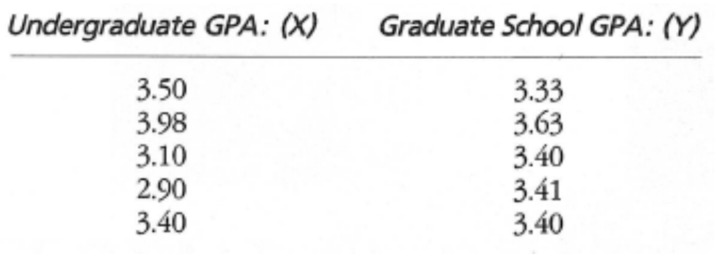
7.1 What is the slope b ?

7.2 What is the intercept a ?

7.3 What is the residual standard deviation s_e ?

7.4 What income would you predict for somebody with 10 years of education?

8. Entrance committee needs to estimate whether a student will be able to gain sufficient grades in first year of college. The studnets need to have 3.00 GPA (average grades, the higher number the better). The committee has data from previous years which include students' GPA in high school and their GPA in the first year of college:



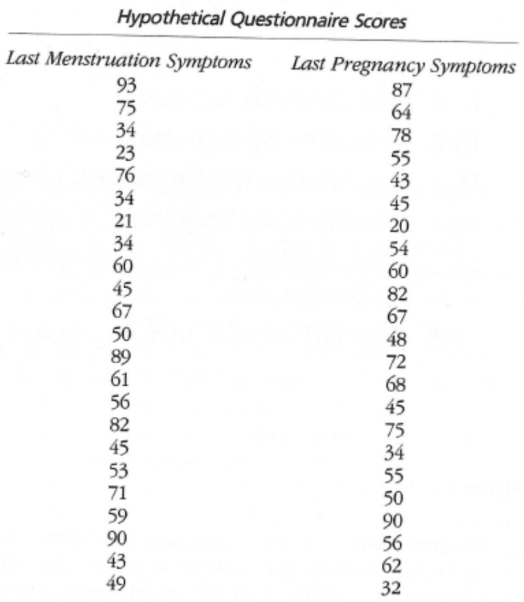
Undergraduate GPA: (X)	*Graduate School GPA: (Y)*
3.50	3.33
3.98	3.63
3.10	3.40
2.90	3.41
3.40	3.40

8.1 What is the slope b ?

8.2 Can you accept student with high school GPA 3.00?

8.3 Between what values of GPA in the first year of college should lie 68% of students who had in high school GPA 3.67?

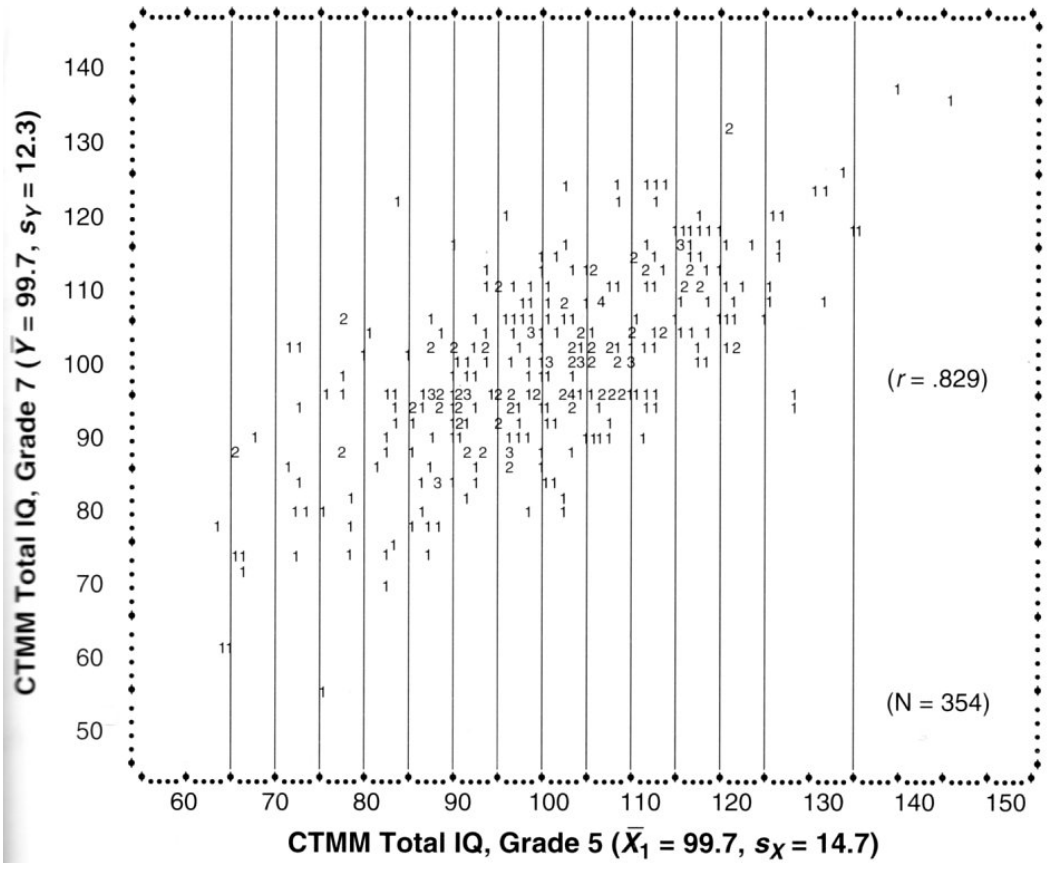
9.2. Carrie (1981) was measuring association between reporting symptoms during pregnancy and during menstruation and association of these reports with general tendency to report psychological and physiological symptoms. Among other results she discovered significant correlation between the number of symptoms experience during menstruation and the number of symptoms experienced during pregnancy. For a woman, who reported 54 menstrual symptoms, predict how many symptom will she report during pregnancy? Identify two values of pregnancy symptoms between which lie 68% of women who reported 54 menstrual symptoms.



Hypothetical Questionnaire Scores

Last Menstruation Symptoms	*Last Pregnancy Symptoms*
93	87
75	64
34	78
23	55
76	43
34	45
21	20
34	54
60	60
45	82
67	67
50	48
89	72
61	68
56	45
82	75
45	34
53	55
71	50
59	90
90	56
43	62
49	32

10. The following graph is scatterplot generated for IQ scores of 354 children tested in 5th grade (X) and 7th grade (Y).



10.1 What is the highest and the lowest score in 5th grade?

10.2 What is the highest and the lowest score in 7th grade?

10.3 Does the association look linear??

10.4 Based on the scatterplot, is the assumption of homoscedascity satisfied?

10.5 Compute slope b.

10.6 Compute intercept a.

10.7 Write down regression equation with the computed coefficients.

10.8 Thomas had in 5th grade IQ score 140. Predict his score in 7th grade.

10.9 David had in 5th grade IQ score 70. Predict his score in 7th grade.

10.10 Draw the regression line in the scatterplot.

10.11 Compute residual standard deviation s_e (s_{res}).

10.12 What percentage of predicted scores will differ from the real values by less than 7 points?

10.13 With approximately 2/3 probability (i.e. 68%) Thomas' IQ in 7th grade lies between ____ and ____ and David's between ____ and ____.

11. A test of intelligence correlates with reading test 0.82. Miq = 100, SDiq = 15, Mr = 8, SDr = 2 and both variable are normally distributed.

11.1 Determine regression equation for prediction of reading (Y) from IQ scores (X).

11.2 What is average reading performance for IQ=100?

11.3 What is average reading performance for IQ=90?

11.4 Compare percentile equivalents of scores X and Y' from the previous question.

11.5 What percentage of children with IQ=90 will have above average reading score? (compute with use of residual standard deviation)

12. In a study investigating internet chatting we read that the relationship between chatting and depression level can be expressed by regression equation $\hat{y} = 0{,}5x - 0{,}2$, where y represents chatting (number of hours spent by chatting weekly) and x represents depression level. The study also shows basic descriptive statistics for both variables:

	N	m	s
chatting	2380	0,6	2,0
depression	2380	1,6	0,8

12.1 What percentage of chatting variance can be explained by depression?

12.2 If we conversely wanted to predict depression from chatting, how would the regression equation look like?

12.3 What depression score would you predict to a person who spends 10 hours per week by chatting?

15. In a study of noise tolerance, a researcher is interested in prediction of time for which people can tolerate very loud annoying sound. He hypothesizes that the more loudly an adolescent listens to his mp3 player, the more time he will be able to tolerate the loud sound. The following table summarizes his results:

Player volume **[*% of maximum volume*]**	**Time of sound tolerance [*s*]**
25	5
31	9
55	20
42	13
47	18
m = 40 *s = 12*	*m = 13* *s = 6*

15.1 Create scatterplot with player volume as predictor.

15.2 Pearson's correlation between the two variables is 0.98. What will be the values of Spearman and Kendall coefficient?

15.3 Determine regression equation for prediction of sound tolerance time from player volume.

15.4 What is residual standard deviation?

15.5 What sound tolerance time would you predict to someone who listens to his mp3 player on 60% volume?

19. In study about creativity the researcher wanted to examine how visual memory influences creativity. They administered a test of visual memory in which the scores express how many from 20 objects can respondent remember after 10 minutes. Creativity was measured on interval scale with *M*=40 and *SD*=10 and visual memory scale had *M*=13 and *SD*=6. Correlation between creativity and visual memory was 0.6.

19.1 By linear regression, estimate creativity score for a person who had visual memory score 8.

19.2 What is approximately the probability that our estimate will differ from the real value by more than 4 points?

19.3 How do we check the assumption of homoscedascity?

20. Which two basic approaches can we choose, if we want to use linear regression, but if the relationship between our variables is not linear?

22. Regression of tolerance on age is expressed by this equation: *Y*=0,22*X*+15,6. This means that if ……………… increases by 10 units, we will estimate for the person by ………… units higher score in ……………. (fill in the answers).

23. What are the consequence, if the assumption of homoscedascity is violated?

28. Mx=3, My=7. Regression line, which predicts Y from X, will intercept (go through) the point (3;7). Is this statement true?

29. The point in which regression line intercepts axis Y is:
a) always mean score of predictor
b) sometime positive value, sometimes negative value
c) always zero

30. Regression line slope:
a) is the same as correlation coefficient
b) can be positive or negative
c) both options are true