Data Analytics
(CS40003)

Practice Set VI

(Topic: Regression Analysis)

I. Concept Questions

1. The need for a nonlinear regression can only be determined by a lack of fit test.

2. The correlation coefficient indicates the change in y associated with a unit change in x.

3. To conduct a valid regression analysis, both x and y must be approximately normally distributed.

4. Rejecting the null hypothesis of no linear regression implies that changes in x cause changes in y.

5. In linear regression we may extrapolate without danger.

6. If x and y are uncorrelated in the population, the expected value of the estimated linear regression coefficient (slope) is zero.

7. If the true regression of y on x is curvilinear, a linear regression still provides a good approximation to that relationship.

8. The x values must be randomly selected in order to use a regression analysis.

9. The error or residual sum of squares is the numerator portion of the formula for the variance of y about the regression line.

10. The term $\hat{\mu}_{y|x}$ serves as the point estimate for estimating both the mean and individual prediction of y for a given x.

11. Useful prediction intervals for y can be obtained from a regression analysis.

12. In a regression analysis, the estimated mean of the distribution of y is the sample mean ($\bar{y}$).

13. All data points will fit the regression line exactly if the sample correlation is either +1 or −1.

14. Given that SSR = 50 and SSE = 100, calculate R^2.

15. The multiple correlation coefficient can be calculated as the simple correlation between and …

16. (a) What value of R2 is required so that a regression with five independent variables is significant if there are 30 observations? [Hint: Use the 0.05 critical value for F(5, 24)].

    (b) Answer part (a) if there are 500 observations.

    (c) What do these results tell us about the R^2 statistic?

17. If x is the number of inches and y is the number of pounds, what is the unit of measure of the regression coefficient?

18. What is the common feature of most “influence” statistics?

19. Under what conditions are least squares not the best method for estimating regression coefficients?

20. What is the interpretation of the regression coefficient when using logarithms of all variables?

21. What are the basic principle underlying inferences on partial regression coefficients?

22. Why is multicollinearity a problem?

23. List some reasons why variable selection is not always an appropriate remedial method when multicollinearity exists.

24. In multiple regression, the coefficient R can be interpreted as
    a. the percentage of variance accounted for in the dependent variable by the set of independent variables.
    b. the percentage of variance accounted for in the dependent variable by a single independent variable.
    c. the strength of a relationship between the dependent variable and a set of independent variables
    d. all of the above.

25. In multiple regression, the R Square can be interpreted as
    a. the percentage of variance accounted for in the dependent variable by the set of independent variables.
    b. the percentage of variance accounted for in the dependent variable by a single independent variable.
    c. the strength of a relationship between the dependent variable and a set of independent variables
    d. the percentage of variance accounted for in the dependent variable by the set of independent variables minus an estimate penalty.

26. In multiple regression, the Adjusted R Square can be interpreted as
    a. the percentage of variance accounted for in the dependent variable by the set of independent variables.
    b. the percentage of variance accounted for in the dependent variable by a single independent variable.
    c. the strength of the relationship between the dependent variable and the set of independent variables.
    d. the percentage of variance accounted for in the dependent variable by the set of independent variables minus an estimate penalty.

27. In a multiple regression analysis, the final section of the output contains the coefficients. Which of these coefficients is of primary concern?
    a. unstandardized B
    b. standard error of B
    c. standardized coefficient beta
    d. standard error of beta

28. The prediction interval for y is widest when x is at its mean.

Oxidation	Temperature
4	-2
3	-2
3	0
2	1
2	2

The data of the above Table represent the thickness of oxidation on a metal alloy for different settings of temperature in a curing oven. The values of temperature have been coded so that zero is the "normal" temperature, which makes manual computation easier.

    a. Calculate the estimated regression line to predict oxidation based on temperature. Explain the meaning of the coefficients and the variance of residuals.
    b. Calculate the estimated oxidation thickness for each of the temperatures in the experiment.

c. Calculate the residuals and make a residual plot. Discuss the distribution of residuals.
d. Test the hypothesis that $\beta 1 = 0$, using both the analysis of variance and t tests.

29.

Days	Sugar
0	7.9
1	12.0
3	39.5
4	11.3
5	11.8
6	11.3
7	4.2
8	0.4

The data of above Table show the sugar content of a fruit (Sugar) for different numbers of days after picking (Days).

a. Obtain the estimated regression line to predict sugar content based on the number of days the fruit is left on the tree.
b. Calculate and plot the residuals against days. Do the residuals suggest a fault in the model?

30.

Midterm	Final
82	76
73	83
95	89
66	76
84	79
89	73
51	62
82	89
75	77
90	85
60	48
81	69
34	51
49	25
87	74

The grades for 15 students on midterm and final examinations in an

English course are given in Table.

a. Obtain the least-squares regression to predict the score on the final examination from the midterm examination score. Test for significance of the regression and interpret the results.
b. It is suggested that if the regression is significant, there is no need to have a final examination. Comment. (Hint: Compute one or two 95% prediction intervals.)
c. Plot the estimated line and the actual data points. Comment on these results.
d. Predict the final score for a student who made a score of 82 on the midterm. Check this calculation with the plot made in part (c).
e. Compute r and r2 and compare results with the partitioning of sums of squares in part (a).

31.

City	State	Lat	Range	City	State	Lat	Range
Montgome	AL	32.3	18.6	Tuscon	AZ	32.1	19.7
Bishop	CA	37.4	21.9	Eureka	CA	40.8	5.4
San Dieg	CA	32.7	9.0	San Fran	CA	37.6	8.7
Denver	CO	39.8	24.0	Washington	DC	39.0	24.0
Miami	FL	25.8	8.7	Talahass	FL	30.4	15.9
Tampa	FL	28.0	12.1	Atlanta	GA	33.6	19.8
Boise	ID	43.6	25.3	Moline	IL	41.4	29.4
Ft wayne	IN	41.0	26.5	Topeka	KS	39.1	27.9
Louisv	KY	38.2	24.2	New Orl	LA	30.0	16.1
Caribou	ME	46.9	30.1	Portland	ME	43.6	25.8
Alpena	MI	45.1	26.5	St cloud	MN	45.6	34.0
Jackson	MS	32.3	19.2	St Louis	MO	38.8	26.3
Billings	MT	45.8	27.7	N PLatte	NB	41.1	28.3
L Vegas	NV	36.1	25.2	Albuquer	NM	35.0	24.1
Buffalo	NY	42.9	25.8	NYC	NY	40.6	24.2
C Hatter	NC	35.3	18.2	Bismark	ND	46.8	34.8
Eugene	OR	44.1	15.3	Charestn	SC	32.9	17.6
Huron	SD	44.4	34.0	Knoxvlle	TN	35.8	22.9
Memphis	TN	35.0	22.9	Amarillo	TX	35.2	23.7
Brownsvl	TX	25.9	13.4	Dallas	TX	32.8	22.3
SLCity	UT	40.8	27.0	Roanoke	VA	37.3	21.6
Seattle	WA	47.4	14.7	Grn bay	WI	44.5	29.9
Casper	WY	42.9	26.6				

The above Table gives latitudes (Lat) and the mean monthly range (Range) between mean monthly maximum and minimum temperatures for a selected set of U.S. cities.

a. Perform a regression using Range as the dependent and L at as the independent variable. Does the resulting regression make sense? Explain.

b. Compute the residuals; find the largest positive and negative residuals. Do these residuals suggest a pattern? Describe a phenomenon that may explain these residuals

32. The thrust of an engine (y) is a function of exhaust temperature (x) in ◦F when other important variables are held constant. Consider the following data.

x	y	x	y
4300	1760	4010	1665
4650	1652	3810	1550
3200	1485	4500	1700
3150	1390	3008	1270
4950	1820		

a. Plot the data.
b. Fit a simple linear regression to the data and plot the line through the data.

33. A set of experimental runs was made to determine a way of predicting cooking time y at various values of oven width x1 and flue temperature x2. The coded data were recorded as follows:

y	**x1**	**x2**
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

Estimate the multiple linear regression equation

$\mu_Y|x1,x2 = \beta0 + \beta1x1 + \beta2x2.$

34. An experiment was conducted to determine if the weight of an animal can be predicted after a given period of time on the basis of the initial weight of the animal and the amount of feed that was eaten. The following data, measured in kilograms, were recorded:

Final Weight, y	Initial Weight, x1	Feed Weight, x2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

a. Fit a multiple regression equation of the form

$\mu_Y|x1,x2 = \beta0 + \beta1x1 + \beta2x2.$

b. Predict the final weight of an animal having an initial weight of 35 kilograms that is given 250 kilograms of feed.

35. An experiment was conducted on a new model of a particular make of automobile to determine the stopping distance at various speeds. The following data were recorded.

Speed, v (km/hr)	35	50	65	80	95	110
Stopping Distance, d (m)	16	26	41	62	88	119

a. Fit a multiple regression curve of the form
$\mu_{D|v} = \beta0 + \beta1v + \beta2v2.$
b. Estimate the stopping distance when the car is traveling at 70 kilometres per hour.

36. The following data are given:

x	0	1	2	3	4	5	6
y	1	4	5	3	2	3	4

a. Fit the cubic model $\mu_Y|x = \beta0+\beta1x+\beta2x2+\beta3x3.$
b. Predict Y when x = 2.

37. An experiment was conducted in order to determine if cerebral blood flow in human beings can be predicted from arterial oxygen tension

(millimetres of mercury). Fifteen patients participated in the study, and the following data were collected

Blood Flow, y	Arterial Oxygen Tension , x
84.33	603.40
87.80	582.50
82.20	556.20
78.21	594.60
78.44	558.90
80.01	575.20
83.53	580.10
79.46	451.20
75.22	404.00
76.58	484.00
77.90	452.40
78.80	448.40
80.67	334.80
86.60	320.30
78.20	350.30

Estimate the quadratic regression equation

$$\mu_{Y|x} = \beta 0 + \beta 1x + \beta 2x2.$$