GEOG 413/613

LECTURE 7

1

Regression Analysis

- To understand relationship between variables:
  - we need a measure: correlation
  - correlation indicates the extent and sign of relation
  - to prove if the measure is statistically reliable.

  - No functional/causal relationship is assumed between the two variables

2

2

Regression Analysis

- Correlation does not reflect the nature of relationship between the variables
- If we find a significant correlation between variables, this could mean that A depends on B, B depends on A, A and B depend on each other, or A and B depend on a third variable C but have no relation to each other.

Regression Analysis

- A famous example is the correlation between ice cream sales and home fires.
  - It would be strange to suggest that eating ice cream causes people to start fires, or that experiencing fires causes people to buy ice cream.
  - In fact, both of these parameters depend on air temperature.

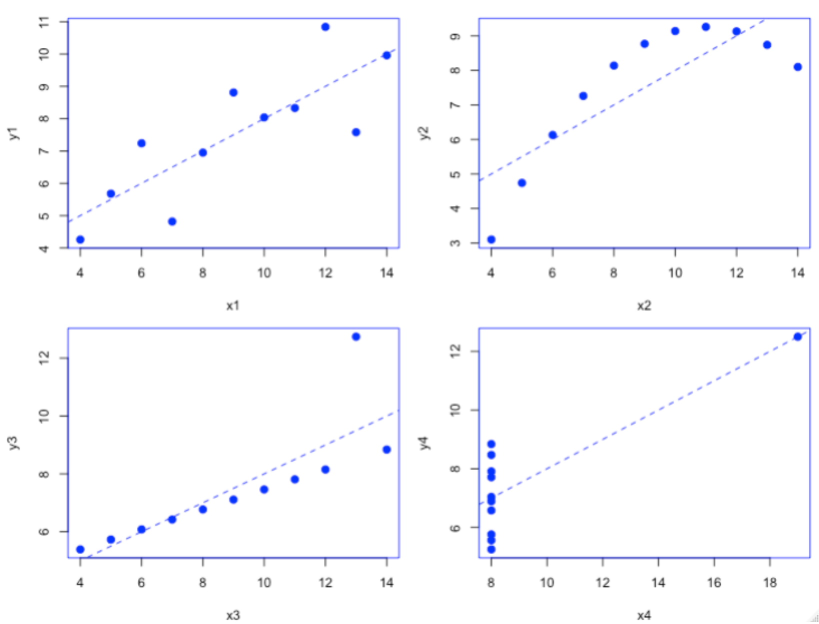
Regression Analysis

- An important example of relationships where numbers alone do to provide a reliable answer, is the *Anscombe's quartet*
  - four sets of two variables which have almost identical means and standard deviations, however their scatter plots are remarkably different

	x1	x2	x3	x4	y1	y2	y3	y4
mean	9	9	9	9	7.500909	7.500909	7.50000	7.500909
var	11	11	11	11	4.127269	4.127629	4.12262	4.123249

5

5



6

6

Regression Analysis

- To measure the extent and sign of linear relationship, we need to calculate *correlation coefficient*.
- The absolute value of the correlation coefficient varies from 0 to 1.
- Zero means that the values of one variable are unconnected with the values of the other variable.
- A correlation coefficient of 1 or −1 is an evidence of a linear relationship between two variables.
- A positive value of means the correlation is positive (the higher the value of one variable, the higher the value of the other), while negative values mean the correlation is negative (the higher the value of one, the lower of the other).
  - **QN: If the Correlation Coefficient is Zero, how would the scatter of the variables look like?**

7

7

Regression Analysis / Bivariate Analysis

- Regression Analysis is sometimes referred to as Bivariate Analysis if two two variables are explored in detail
  - Assumption is that one variable influences/affects the other
  - Example:
    - The relationship between precipitation and population density.
    - The assumption is the amount of moisture at locations influences population density

8

8

Bivariate Regression

- Similar to correlation analysis, bivariate regression seeks to examine the influence of one variable on another
- Independent Variable (Explanatory Variable)
  - The variable creating the influence/effect
- Dependent Variable
  - The variable receiving the influence or effect

9

9

Relationships in Bivariate Regression

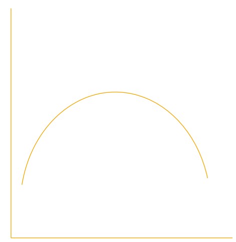
- In correlation, the assignment of the axes for the variables is arbitrary
- In binary regression
  - Independent variable on the X-axis
  - Dependent variable on the Y-axis
- The form of association between the variables can be portrayed using a scatterplot

10

10

Relationships in Bivariate Regression

- Not Significant (Statistically)
- Linear (Positive, Negative)
- Curve-linear (Concave, Convex)



Concave

Convex

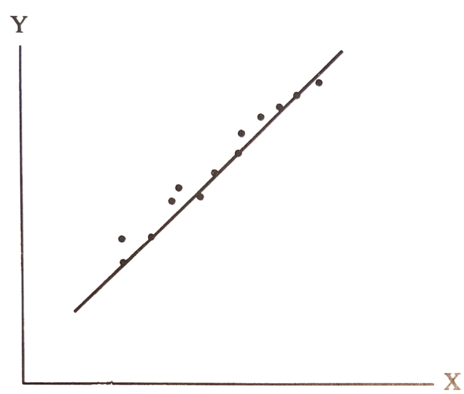
- Undefined complex (statistically significant but relationship cannot be reliably described)

11

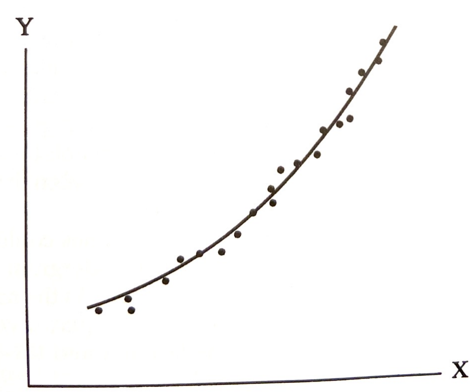
11

Relationships

Case 1:
Linear



Case 2:
Curvi-linear



12

12

Bivariate Regression Line

- The point pattern on the scatterplot can can be described with a least-squares regression line
- The least-squares regression line is unique
  - Minimizes the sum of the squared vertical distances between each data point and the line

13

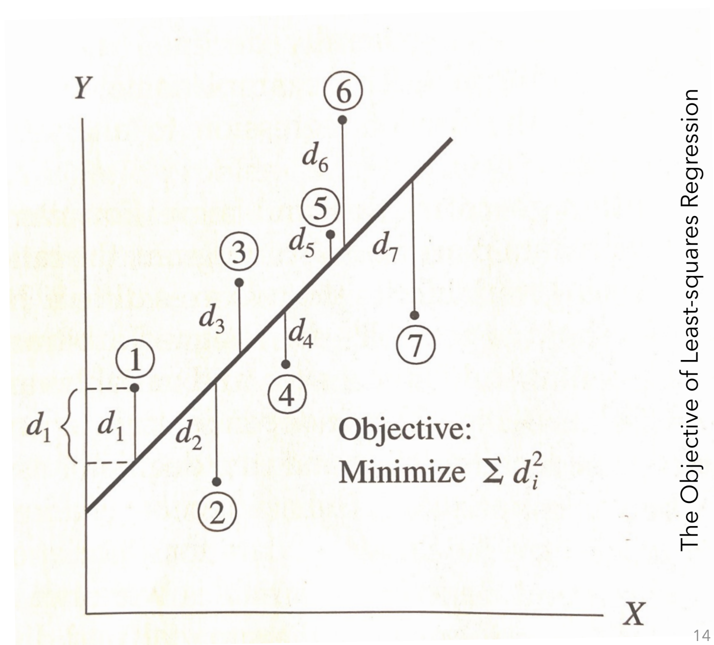
13

Bivariate Regression Line

- The regression is given by the straight-line equation

$$Y = a + bX$$

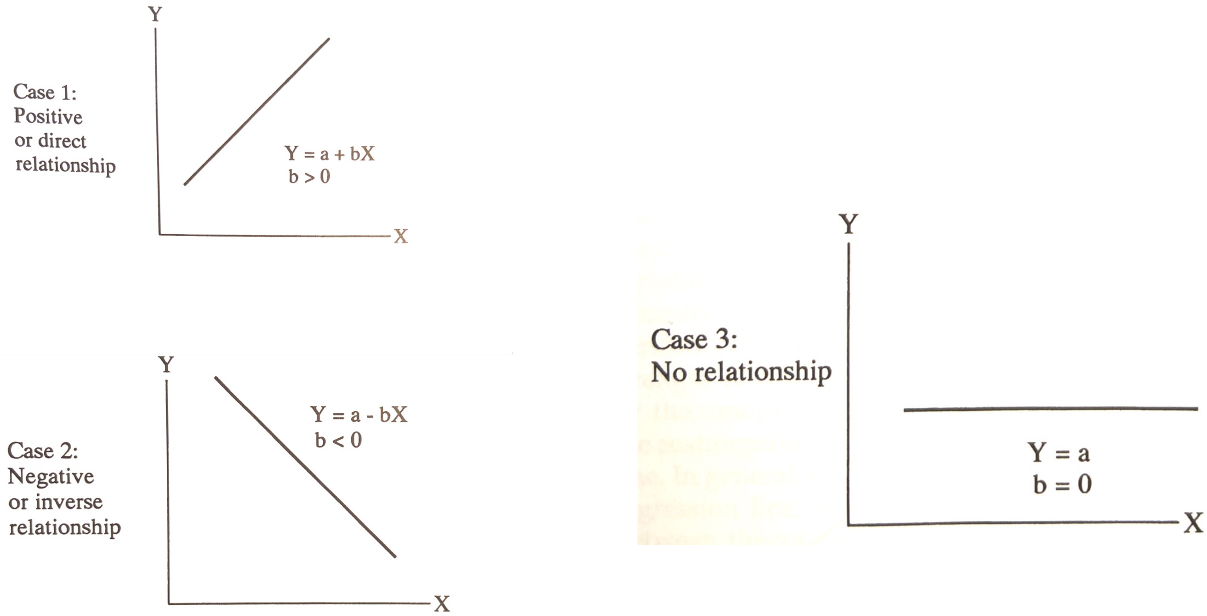
- *a* is the intercept on the Y-Axis
  - Represents the value of Y when X is zero
- *b* represents the slope of the line
  - Also, the correlation coefficient



The Objective of Least-squares Regression

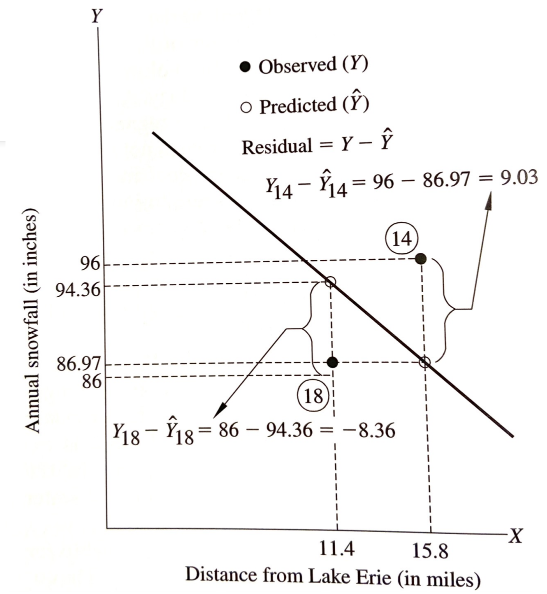
14

14



Strength of Relationship

- The ability of the independent to account for variation in Y provides a measure of the strength of the relationship
- The closer the points to the regression line stronger the relation between the variables
- Strength is measured with the coefficient of determination, r^2
- r^2 = ratio of explained variation to total variation



17

17

Variation in the dependent variable

- Variation in the dependent has two parts
  - Explained Variable
  - Unexplained variable

$$\sum y^2 = \sum y_e^2 + \sum y_u^2$$

$$\sum y^2 \Rightarrow TSS\ (Total\ Sum\ of\ Squares), total\ variation\ in\ Y$$

$$\sum y_e^2 \Rightarrow explained\ variation\ (caused\ by\ independent\ variable)$$

$$\sum y_u^2 \Rightarrow unexplained\ variation$$

18

18

The Coefficient of Determination

- Some phenomena many be modeled by the regression line well, others not.

- Coefficient of determination comes in handy
- It is the ratio between the predicted values of Y_p (regression variance) and the variance of the observed Y_o
- Suppose the C.D. is 0.625, then we can say that 62.5% of the dependent variable is accounted for in the predicated values (Y_p, the regression line, the predicted variable)

19

19

The Coefficient of Determination

$$\sum y_e^2 = \frac{(\sum xy)^2}{\sum x^2}$$

$$\sum x^2 \Rightarrow total\ variation\ of\ X$$

$$(\sum xy)^2 \Rightarrow the\ square\ of\ the\ Covariation\ of\ X\ and\ Y$$

$$r^2 = \frac{\sum y_e^2}{\sum y^2}$$

r^2 = coefficient of determination

20

20

Demo

- Understanding the Variation in the independent variable
  - Create a variable X, where x=x+2
  - Create a variable Y_1, where y_1=x*1.5
  - Create a variable Y_2, where $y_2 = y_2$ + random number between -2 and +2
  - Y_1 – total variation all explained by X
  - Y_2 – total variation explained in part by X and some unknown variance

21

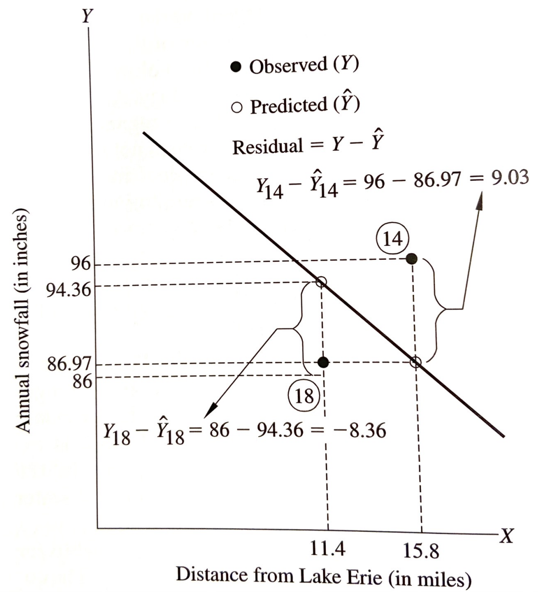
21

X	Y1	Y2
2	3	5
4	6	8
6	9	8
8	12	11
10	15	13
12	18	19
14	21	23
16	24	23
18	27	26
20	30	32
22	33	31
24	36	38
26	39	39
28	42	40
30	45	46
32	48	48
34	51	50
36	54	54
38	57	58
40	60	60
42	63	61
44	66	64
46	69	71
48	72	72
50	75	75
52	78	79
54	81	83
56	84	85
58	87	87

Var X	280		
Var Y1	630		
Var Y2	634.256837		
Covariance (X&Y1)	420		
Covariance (X&Y2)	420.758621		
Exp Var Y1	(420*420)/280 = 630		
Exp Var Y2	??		

22

22



23

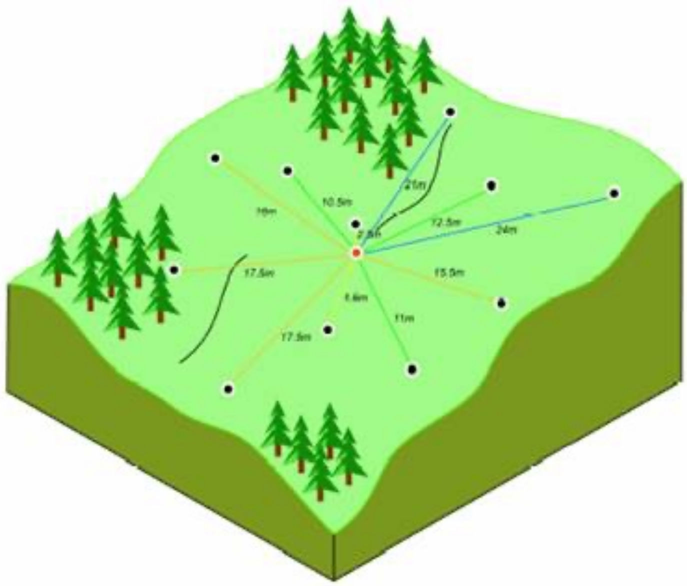
23

Semivariance

- In the context of spatial points on a surface there is a need to understand the degree of association between points
- Semivariance used in the process of Kriging
  - Equal to half the variance of the differences between all possible points spaced a constant distance apart
  - At d=0 semivariance is Zero, as the points further are considered, the semivariance increases until it reaches the variance of the whole surface.
    - This is is the maximum distance at which two points are related
    - This maximum distance is called the range
    - The range defines the size of the neighborhood over which control points should be selected to predict other points

24

24



25

25