ومن يؤت الحكمة
فقد اوتي خيرا كثيرا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

Assignment 1

By

SAJJAD ULLAH

M.Phil 2nd Semester (Regular)

Submitted To

Dr. ZAHID ASGHAR

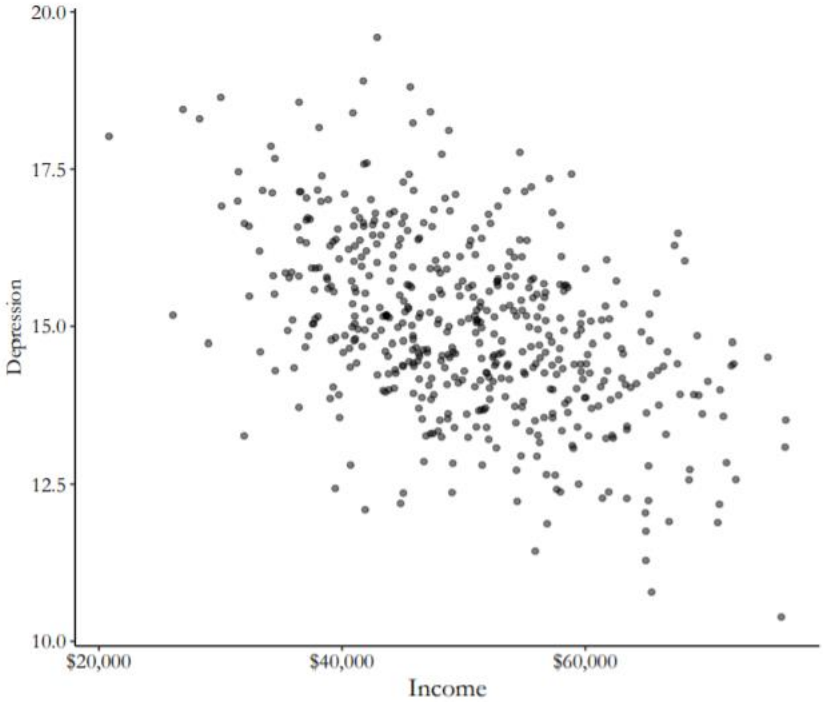
School of Economics

Quaid-i- Azam University Islamabad

1. What is a conditional distribution?

Answer: If X and Y are two jointly distributed random variables, and then the conditional distribution of Y given X is the probability distribution of Y when X is known to be a certain value. If we want to know the probability that a person prefers a certain sport given that they are male, then this is an example of a conditional distribution. The value of one random variable is known (the person is male), but the value of the other random variable is unknown.

2. The following figure (using fictional data) describes the relationship between Income level and rating on a scale testing for signs of Depression.



a. How does the conditional mean of Depression change as Income increases?

Answer: The scatterplot shows the negative relationship between the income and depression. As the income increases the level of depression decreases. When an individual has sufficient amount of income so his/her needs fulfill more and more so he/she will be free from depression.

b. Does the graph indicate that lower income causes depression? Why or why not?

Answer: Yes, the graph indicates lower income causes depression because there is inverse relation between income and depression.

3. The below fictional table depicts data collected from 3000 university students on their classification (Freshman, Sophomore, Junior, Senior) and whether or not they receive financial aid. The table shows a cross tabulation of classification and receipt of financial aid.

Financial Aid	Freshman	Sophomore	Junior	Senior
Yes	508	349	425	288
No	371	337	384	338

a. Calculate the probability of receiving financial aid given that a student is a Senior.

Answer

Probability of senior that receive financial aid can be calculated by the following formula as

Prob(senior) = Senior/total number of students.

Prob(senior) = 288/(508+371+349+337+425+384+288+338)

Prob(senior) = 288/(3000)

Prob(senior) = 0.096

b. Calculate the probability that a student is a Senior given that they receive financial aid.

Prob(senior) = Senior/total number of students they receive financial aid.

Prob(senior) = 288/(508+349+425+288)

Prob(senior) = 288/(1570)

Prob(senior) = 0.183

c. Calculate the probability of receiving financial aid given that a student is a Freshman.

Prob(freshman) = Senior/total number of students.

Prob(freshman) = 508/(3000)

Prob(freshman) = 0.169

4. Describe two advantages and one disadvantage of using line-fitting methods as opposed to calculating local means.

Advantages

The line of best fit is used to express a relationship in a scatter plot of different data points.

The advantage of using the line-fitting method as opposed to local means is that the line is estimated using all the data than rather using the local data only, due to which the results are precise. Furthermore, we can add more than one variable.

Disadvantage

Disadvantage associated with using line-fitting method is that we need to fit a line that is we have to pick the shape for the relationship; it will get the best shape we give it. However, if the shape we selected is wrong the estimate of conditional mean will be all wrong.

5. Consider the line of best fit: $Y = 4 - 3.5X$.

a. What is the conditional predicted mean of Y when $X = 6$?

Solution: Putting the value of X = 6 in the main equation:

Y = 4 – 3.5 (6)

Y = 4 – 21

Y= -17

Her the conditional predicted mean from the above eqation is -17 when X = 6.

b. What is the conditional predicted mean of Y when$X = -2$?

Solution

Putting the value of X = -2 in the main equation:

Y = 4 – 3.5 (-2)

Y = 4 + 7

Y = 11

Her the conditional predicted mean from the above eqation is -17 when X = 6.

6. Which of the following terms describes a measurement of how much two variables vary with each other?

a. Variance
b. Conditional mean
c. Covariance
d. Local mean

Answer is Covariance

7. What is the difference between covariance and correlation?

Covariance

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or, we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. It is a measure to show the extent to which given two random variables change with respect to each other. Covariance can have both positive and negative values. Based on this, it has two types. Positive Covariance and negative Covariance

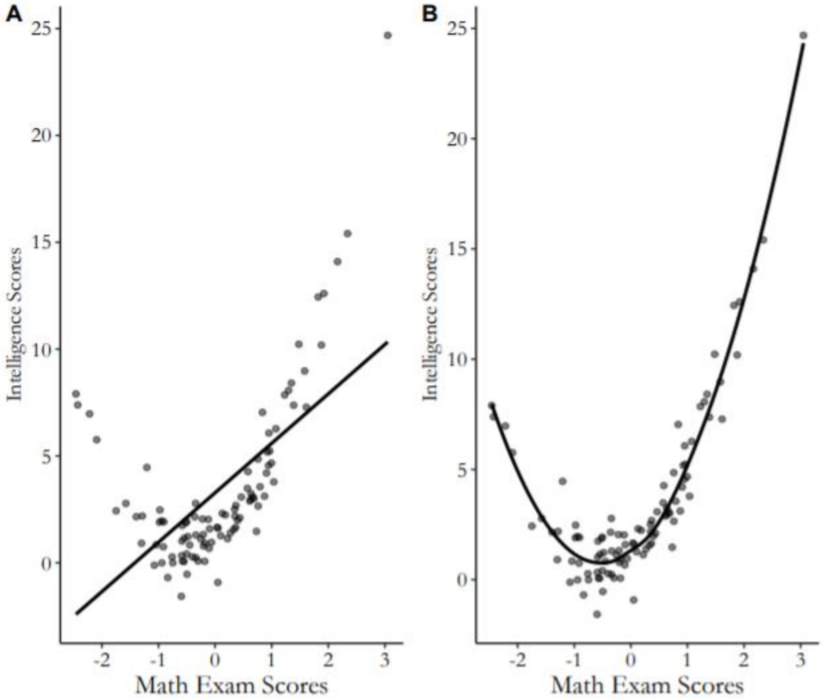
- ✓ It is a measure of correlation.
- ✓ The value of covariance lies between -∞ and +∞.
- ✓ It indicates the direction of the linear relationship between the given two variables.

Correlation

- ✓ It is a measure used to describe how strongly the given two random variables are related to each other.
- ✓ It is defined as the scaled form of covariance.
- ✓ The value of correlation lies between -1 and +1.
- ✓ It measures the direction and strength of the linear relationship between the given two variables.

8. Figure A and Figure B below depict the (fictional-data) relationship between scores on a math exam and an intelligence measure from data collected from a fictional sample of 100

students.



a. What type of shape is fitted in Figure A?   Answer is linear.
b. What kind of shape is fitted in Figure B?   Answer is upper parabola.
c. Which shape is a better fit for the data, and how can you tell?

Answer: The line in figure B is best fitted, because it incorporates the conditional means more accurately than that in the figure A.

d. For Figure A, describe the residuals for different ranges of math exam scores. Are the observed data below or above the line/curve? Are the residuals evenly scattered around the line/curve?

Answer: On the bases of point 2 the residuals are above the estimated line, on point -1 the residuals are below the estimated line up to point 1. Between point 1 and 2 some values are below and others are above the line, from point 2 to 3 the residuals are above the line. The observed data as a whole is neither above nor below the line it is somewhat in between. No, the residuals are not evenly scattered around the line from the range -2 to -1 the values are way above the line, from -1 to 1 the values are below the line but are close to the line unlike the previous one. And from 1 to 3 again the values are above the line and the distance between the residuals and the line is more than the ones below the line.

9. The below table contains fictional data on 5 employees from a company, repotting on how well they get along with their coworkers (GetAlong) and their level of job satisfaction (Satisfaction). The Prediction variable is the predicted satisfaction level, or the conditional mean of satisfaction, for each value of GetAlong derived after fitting a line of best fit using ordinary least squares estimation.

GetAlong	Satisfaction	Prediction	Residual
4.7	5.07	4.72	-0.02
4.21	4.05	4.28	-0.07
5.42	5.33	5.38	0.04
4.14	4.02	4.22	-0.08
3.3	3.59	3.45	-0.15

a. Fill in the "residual" column.

b. Describe how ordinary least squares uses residuals when fitting a line.

Answer: The Ordinary least square uses the residuals when fitting a line in the way, it takes the square of residuals and then sum up all the residual squared residuals. Which is very useful while fitting a line because by this method we will get to know what the best fit for the line is. Without this method there no such method to go for the best fitted line. In our example the sum of square residuals is 0.2375 which is point where the line will pass and is the best fit in our case.

10. We'll be thinking here about the process of controlling for a variable. Consider the example: What is the relationship between first generation status and graduation rate in a population of college students?

a. Give an example of at a variable that might explain why first generation status and graduation rate are correlated other than one causing the other.

Answer: The one variable that I can think of might be parents' education.

b. Describe in five steps how you would subtract out the part of the first-generation/graduation-rate relationship that is explained by the variable you listed in part a.

- ➢ We will get the mean of first-generation status conditional on family income.

- Will subtract this mean from the actual value of first generation and will residual of first generation.
- Will get the mean of graduation rate conditional on family income.
- Will subtract the mean from actual value of graduation rate and will get the residual of graduation rate.
- They will describe the relationship between residual of first-generation status and the residual of graduation rate.

c. How would you interpret the first-generation/graduation-rate relationship you get after performing the steps in part b?

Answer: The interpretation of the first-generation/graduation-rate relationship would be, the result that we get after performing the steps in part B, are free from the influence of family income that was effect both the variables, now the relationship we get is more precise and the relationship is purely between the two variables that we had at first hand. And there is no autocorrelation problem.