

# 1 ☐ Bivariate Descriptive Statistics

## V. Measures of Association for Interval and Ratio Level Data

### 2 ☐ A. Introduction

- Order of Correlation and Regression
  - Correlation Coefficient is the measure of association
  - Regression equation is used to predict the values of Y based upon the values of X
  - You calculate the regression line/plane first, and then the correlation coefficient
- Lecture: Scattergram, Regression Analysis, Correlation

### 3 ☐ B. Scattergram

- Parallel to the contingency table for interval/ratio level data
- Proper Form

### 4 ☐ Scattergram Cont'd

- What it can tell us
  - Existence of a relationship
  - Direction of a relationship
  - Strength of a relationship
  - Form of a relationship
- Examples:

### 5 ☐ Scattergram: No Relationship

### 6 ☐ Scattergram: Weak Relationship

### 7 ☐ Scattergram: Moderate Relationship

### 8 ☐ Scattergram: Strong Relationship

### 9 ☐ Scattergram: Negative Relationship

### 10 ☐ Scattergram: Curvilinear Relationship

### 11 ☐ C. Prediction

Prediction is a means for assessing the presence and strength of any relationship/covariation that exists between two variables

At the most general level:

- If we have a relationship between two variables, we can use the values of one variable to predict the values of the other
- If there is a relationship, our predictions using X should be better than just guessing or predicting by using a measure of central tendency (e.g., mean)
- The stronger the relationship, the better our predictions --- the less prediction error

Thus, one way to assess the existence and strength of the relationship between variables is to determine IF and HOW WELL we can predict one from the another

### 12 ☐ 1. Establishing a baseline for comparison

- If you are to predict the values of Y without any knowledge of independent variables, the best solution is to predict the mean of Y for all cases
  - This is similar to Lambda, where you predict the modal value for all cases
    - Also remember if the distribution is normal, the mean and the mode are the same value
  - Predicting the mean for all cases will result in prediction error, but less error than if we just guessed or

used any other value

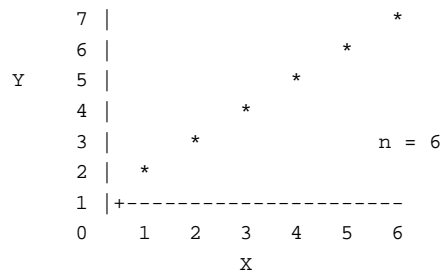
– Why?

- Sum of the deviations about the mean is zero
- Sum of the squared deviations about the mean is smaller than the sum of the squared deviations around any other value in the distribution

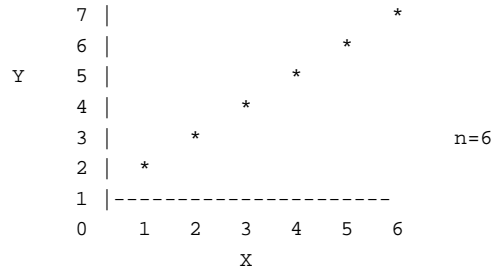
13  2. Summary of Using X to predict Y

- We can use the values of X to predict the values of Y
- If there is a relationship between the variables, the predictions using the values of X should be more accurate than just predicting the grand mean of Y for each case
- Stated otherwise, we will make fewer errors than if we predicted the grand mean for all cases
- OR the stronger the relationship, the fewer prediction errors we will make
- Use regression analysis to predict Y from X, and then use correlation to measure the strength or reduction in prediction error

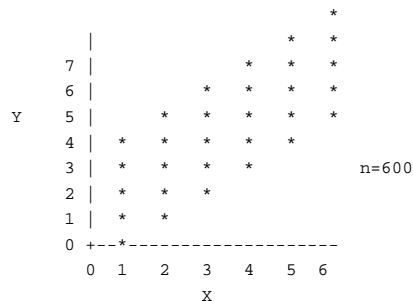
14



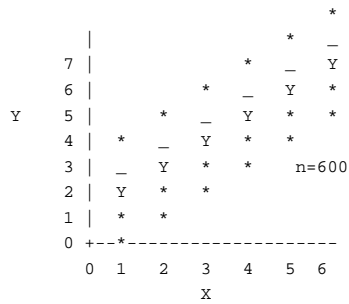
15



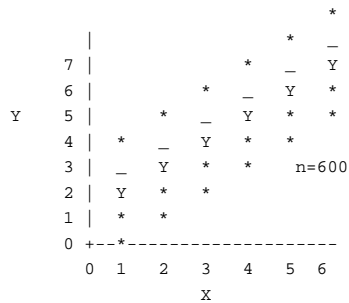
16  Increase the cases to 600 with different values on Y for cases with the same value on X



17  Calculate the mean of the Y for all cases with the same value on X

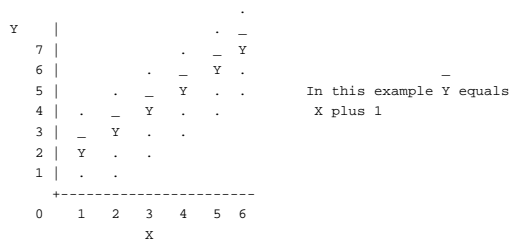


18  Regression Path: the group means on Y for the fixed values of X

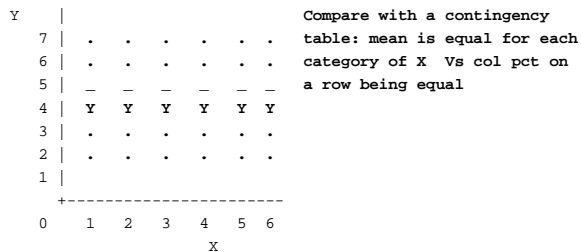


19  Regression Path for a **Positive Relationship**

As X increases so does the average value of Y  
 Low values on X occur with low average values on Y and  
 High values on X occur with high average values on Y



20  Regression Path When There is **No Relationship**



21  Linear Regression Equation for Population

If we can assume

1. Regression path is linear
2. Distribution of Ys on X is normal
3. Variances of all Y distributions on X are identical, *the path of the*

means can be expressed as

$$Y = \alpha + \beta X + e$$

**$\alpha$  is the intercept:**

- Point at which the regression line crosses the Y axis
- Or value of Y when X equals 0

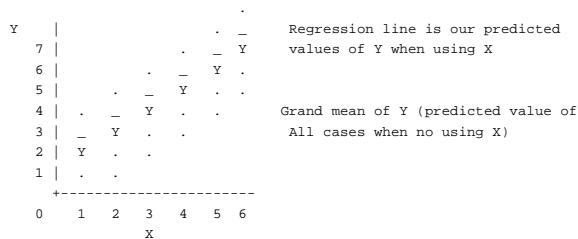
## 22 Linear Regression Equation for Population

## 23 Prediction

- Use the regression equation to predict the value of Y from the value of X
  - Predicted  $Y = \alpha + \beta X$
  - $7 = 1.0 + 1.0(6)$  For those cases where X equals 6
  - $2 = 1.0 + 1.0(1)$  For those cases where X equals 1
- The predicted value of Y is the mean of Y for all cases with this value of X
  - All cases with the same value on X get the same predicted Y
  - We are still using the mean to predict, we have just switched from using the grand mean to using a group mean
  - If there is a relationship, we can get more accurate prediction using these group means instead of the grand mean

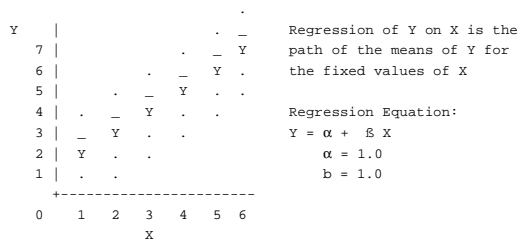
## 24 Prediction Error

Population Regression of Y on X  
A Positive Relationship



## 25 Summary

Population Regression of Y on X  
A Positive Relationship



- Population Vs Sample

## 26 Linear Least Squares Theory

- SAMPLE: Linear Regression line from least squares theory is an estimate of the population regression line
- If we can meet three assumptions, then the regression line established by least squares theory is an unbiased estimate of the population regression line
- Assumptions:
  - 1. Random Sampling
  - 2. Error is random and thus cancels out
  - 3. X and the error are unrelated

## 27

- Least squares estimate of the population regression line  

$$Y = a + bX + e$$

- a is the intercept --- our sample estimate of  $\alpha$
- b is the slope --- our sample estimate of  $\beta$

• **Before examining calculations visualize what LLS has been asked to do**

### 28 Option A: Minimize Deviations

- Visualize what the least squares estimate is trying to do
- It is trying to estimate the line which is the path of the means
  - Its task is to find the unique straight line about which
    - the sum of the deviations is zero
    - the sum of the squared deviations is smaller than the sum of the squared deviations around any other line

### 29 Option B -- Prediction

- Without knowledge of X, the grand mean of Y is the best predictor of Y
  - We predict everyone has the same value
  - The error we make equals the deviations from the mean --which we know sums to zero
- If X & Y are related we can do a better job predicting Y than by just using the grand mean
- So we ask linear least squares to find the line where we make the least prediction error -- is there a line closer to the observed values than the mean of Y??
- If there is no relationship, this estimated line is the grand mean of Y

### 30 No Relationship

- What does LLSQ do if there is no relationship?
  - the line where the sum of the deviations is zero and the sum of the squared deviations from the line is smaller than the sum of the squared deviations about any other line is a horizontal line equal to the grand mean\* of Y  
\*the grand mean is the univariate mean for all cases
  - slope will be zero ( b = 0)  
 intercept will be equal to the grand mean of Y  
 (a = univariate mean of Y)

### 31 Conceptual Computational Formulas

- Calculating a and b

$$a = \bar{Y} - b \bar{X}$$

The SLOPE is Covariation of X and Y divided by the variation in X

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\text{Co-variation of X \& Y}}{\text{Variation of X}}$$

### 32 Numerical Example: Positive

	Values of X	Deviations	Values of Y	Deviations	Covariation
Case	$X_i$	$(X_i - \bar{X})$	$Y_i$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$

IL	5	2	10	2	4
IA	4	1	9	1	1
PA	3	0	8	0	0
IN	2	-1	7	-1	1
MO	1	-2	6	-2	4
$\Sigma$		0		0	10
$\bar{X}$	3				
$\bar{Y}$		8			
$b$					$10/10 = 1.0$

33  Numerical Example: Negative

Case	$X_i$	$(X_i - \bar{X})$	$Y_i$	$(Y_i - \bar{Y})$	
IL	5	2	6	-2	-4
IA	4	1	7	-1	-1
PA	3	0	8	0	0
IN	2	-1	9	1	-1
MO	1	-2	10	2	-4
$\Sigma$		0		0	-10
$\bar{X}$	3				
$\bar{Y}$			8		
$b$					$-10/10 = -1.0$

34  Why “b” does not Measure Strength?

One case example (case IL) to show how variability affects the slope

$$b = \frac{(5 - 3)(10 - 8)}{(5 - 3)^2} = \frac{4}{4} = 1.0$$

Now some one sneaks in and multiplies all the values of X by 10

$$b = \frac{(50 - 30)(10 - 8)}{(50 - 30)^2} = \frac{40}{400} = .1$$

35  Summary on the Slope

- 1. If there is no *linear* relationship between X and Y, the slope will be 0
- 2. If there is no linear relationship between X and Y, the intercept will be the univariate or grand mean of Y
  - Thus, The best way to predict and minimize errors is to predict all cases have the mean value on Y

36  Summary on the Slope

Cont'd

- 3. If the slope does not equal zero, is this because of a real relationship or because of sampling error?
  - Using inferential statistics like student's t, we calculate a probability for the slope: (Prob of getting this non-zero slope if the true population slope was 0)
  - If the probability is less than .05, a relationship is inferred to exist in the population
  - If the probability is greater than .05, we conclude there is no relationship between X and Y

37  Summary on the Slope Cont'd

- **4. If there is a relationship, How strong is it?**

- Because the slope is a function of both the strength of the relationship and the amount of variation, the slope CANNOT be used to indicate strength of the relationship
- This is done by the correlation coefficient and its coefficient of determination

38  **Measure of Association: Correlation Coefficient**

- 1. Use on Interval & Ratio Level Data
- 2. Range is -1.0 to + 1.0 (0 means no linear relationship)
- 3. Symmetrical
- 4. Interpretation: Like Phi -- no desirable interpretation -- use the coefficient of determination instead.

However in the most general sense it measures goodness of fit of the cases to the regression line

39  **Conceptual formula**

Conceptual formula for the correlation coefficient (r)

$$r = \frac{\sum ((X_i - \bar{X})(Y_i - \bar{Y}))}{\sqrt{[\sum(X_i - \bar{X})^2][\sum(Y_i - \bar{Y})^2]}}$$

Co-variation of X & Y  
Square root of the product of variation in X times the variation in Y

40  **The correlation coefficient (r)**

1. r = 0 if there is no *linear* relationship between X and Y

(the slope will also be 0, since both r and the slope have the same numerator)

2. If r is not equal to 0, is this because of a real linear relationship or simple because of sampling error?

If the significance or probability level for r is below .05, we conclude there is a relationship between X and Y

If the significance or probability level for r is above .05, we conclude there is no relationship between X and Y

[The slope and r, will both produce the same decision about the existence of a relationship between X and Y]

41  **r continued**

- 3. If there is a relationship, the sign of r tells the *direction* of the relationship + or -
- 4. If there is a relationship, the size of r tells us the *strength* of the relationship
- 5. While the coefficient of determination provides a more satisfactory interpretation, r can be interpreted as indicating the goodness of fit of the regression line -- how close to the line are the observed values
  - a. If r is 1.0 (the relationship is perfect), all cases will lie on the regression line

- b. As the relationship becomes weaker, the cases are more scattered about the regression line
- c. The weaker the relationship, the less our ability to predict actual values of Y using the regression line -- we have more prediction error

42  **The coefficient of determination ( $r^2$ )**

1. The square of the correlation coefficient has a specific and appealing interpretation

a. *The coefficient of determination is the proportion of total variation in Y accounted for or explained by X*

b. Take the total variation of Y (which is what we want to explain in the first place) and divide into  
 (1) explained variation and  
 (2) unexplained variation

c. The coefficient of determination is the ratio of explained variation to total variation

43  **Comparing Correlation Coefficient and Coefficient of Determination**

$r = .00 \ .30 \ .40 \ .50 \ .70 \ .90 \ 1.00$

$r^2 = .00 \ .09 \ .16 \ .25 \ .49 \ .81 \ 1.00$

Weak      Moderate      Strong

44   **$r^2$  Formulas**

2. Alternative conceptual formulas for the coefficient of determination

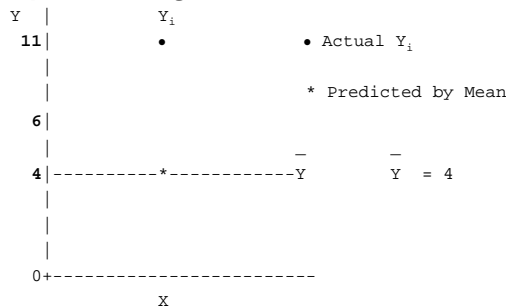
$$r^2 = \frac{\sum (Y_p - Y)^2}{\sum (Y_i - Y)^2} = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{Explained Sums of Squares}}{\text{Total Sums of Squares}}$$

3. What do we mean by explained variation?

Explained variation can be conceptualized as the sum of the squared reductions in original prediction error produced by using the independent variable instead of the mean of Y as the predictor of Y

Unexplained variation can be conceptualized as the sum of the squared residuals, or errors, remaining after Y has been predicted by using X

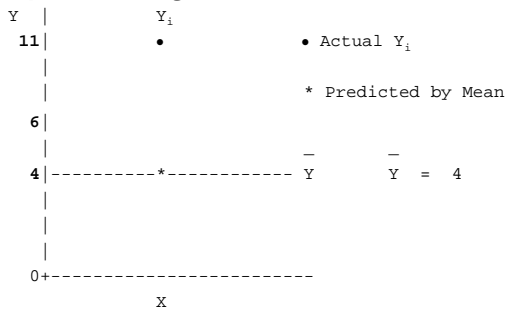
45  **Example Using One Case**



-----For all Cases:

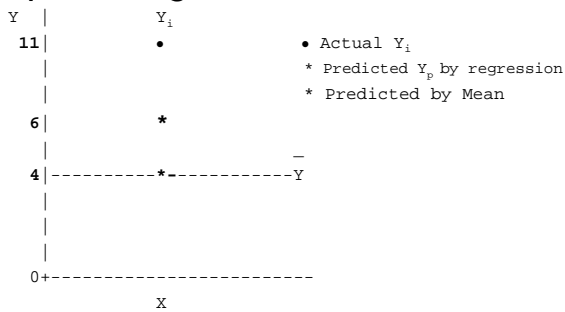
$$\begin{aligned} \text{Total SS} &= \text{Unexplained SS} + \text{Explained SS} \\ \sum (Y_i - Y)^2 &= \sum (Y_i - Y_p)^2 + \sum (Y_p - Y)^2 \end{aligned}$$

#### 46 ☐ Example Using One Case



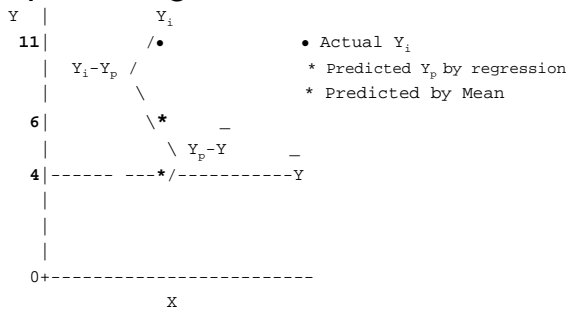
-----For all Cases:  
 Total SS = Unexplained SS + Explained SS  
 $\sum (Y_i - \bar{Y})^2 = \sum (Y_i - Y_p)^2 + \sum (Y_p - \bar{Y})^2$   
 $\sum (Y_i - \bar{Y}) = \sum (Y_i - Y_p) + \sum (Y_p - \bar{Y})$

#### 47 ☐ Example Using One Case



-----For all Cases:  
 Total SS = Unexplained SS + Explained SS  
 $\sum (Y_i - \bar{Y})^2 = \sum (Y_i - Y_p)^2 + \sum (Y_p - \bar{Y})^2$   
 $\sum (Y_i - \bar{Y}) = \sum (Y_i - Y_p) + \sum (Y_p - \bar{Y})$

#### 48 ☐ Example Using One Case



-----For all Cases:  
 Total SS = Unexplained SS + Explained SS  
 $\sum (Y_i - \bar{Y})^2 = \sum (Y_i - Y_p)^2 + \sum (Y_p - \bar{Y})^2$   
 $\sum (Y_i - \bar{Y}) = \sum (Y_i - Y_p) + \sum (Y_p - \bar{Y})$

#### 49 ☐ Conclusion

- Terminology
- Nature of Relationship
- Effect of Outliers on correlation and regression

#### 50 ☐ NonLinear Relationship

- 51  **Outlier: Little Effect**
- 52  **Outlier: Inflates Correlation**
- 53  **Outlier: Smaller Slope & More Accurate Prediction Without the Outlier**
- 54  **Outlier: Creates Negative Relationship**
- 55  **Outlier: Obscures Relationship**