# Chapter 12 Notes: Bivariate Regression & Correlation

## Slope found -

- Correlation - measure of association that tells us how strong & whether or not there is a relationship between 2 variables
- Regression - inferential statistic technique used to see whether or not there is a relationship between 2 vars measured at the interval ratio level
- Regression - prediction equation, we predict the value of the dependent variable when we have the independent variable.

## Different Inferential Stats Technique

- t-test - use if the dependent var is measured at interval-ratio level
- anova - use if the dependent var is measured at interval-ratio level, but only use if you have 3 or more samples, otherwise go with t-test
- Chi-square - use if both your dependent & independent var are nominal and/or ordinal level
- regression - use if both independent var & dependent v ar are measured at the interval-ratio level

## Scatter Diagram

- AKA a scatterplot, is a quick visual method used to display a relationship between 2 interval-ratio variables
- Scatter diagrams/plots are often used as a 1st step in regression analysis to see whether 2 variables are associated

## A Scatterplot

- Typically, the **independent variable** (X) is arranged along the horizontal axis (side to side) and the **dependent variable** (Y) is arranged along the vertical axis (up & down)

## Linear Relationships

- **Linear relationships** - a relaitonship between 2 interval-ratio variables where the observations displayed in a scatter diagram can be approximated with a straight line

(approximating means we can draw a straight line to represent the relationship but it isn't a perfect predictor, NOTE that not all observations will not fall on the straight line).

- **Deterministic (perfect) linear relationship** - a relationship between 2 interval-ratio variables in which all the observations (the dots) fall along a straight line. The line provides a predicted value of Y (the vertical axis) for any value of X (the horizontal axis). NOTE that these are rare to see, it's more of a theoretical concept.

# Equation for a Straight Line

$$Y = a + b(X)$$

- A = Y Intercept (value of Y when X is 0)
- B = slope (aka Rise/Run)
- Y = dependent variable
- X = independent variable

# Finding the fitting line

- Unfortunately, most relationships we study in the social sciences aren't deterministic
- The best fitting line is the line that gets as close as possible to all the observations (dots) in the distribution
- You find the best fitting line when the dependent variable (Y) is not completely determinde by the independent variable (X) so not all the observations will lie exactly on the line
- The best fitting line is the one that produces the least amount of error among the dots in the scatter diagram. It's the line that gets as close as possible to all the dots in the distribution

# Defining Error

- Error is when the actual observed value is not on the line, meaning the observed value is different from the predicted value

There are 2 values of Y:

1. A predicted value of Y that's symbolized as Y (letter Y with "^" on the top, aka y-hat) & which is generated by the prediction equation (AKA the linear regression equation)
2. The observed Y that is symbolized as Y
3. Error is the difference between these 2 Y's

# The Sum of Squared Errors

- We want a line that will minimize error for each observation. However, any line we choose will minimize the errors for some observations & maximize errors for others
- We want to find a prediction equation (aka linear regression equation) that will minimize errors over all observations, thisi the best fitting line

# The Sum of Squared Errors

- There are many mathematical ways of defining error
- It's prefered to use the sum of the absolute errors or sum of squared errors in statistics
- We simply square & sum the errors over all observations

$$\sum e^2 = \sum (Y - \hat{Y})^2$$

1. First you find out the error by subtracting the actual observed value from the prediction
2. Then you square that value & repeat for all observations, then add up them all up

# The Least Squares Line

- The best fitting regression line is the line where the sum of the squared errors is at a minimum
- This line is called the least-squares line & the technique that produces this line is called the least-squares method
- The technique involves choosing *a* & *b* for the equation Y (^) = a + bX such that the squared errors will have the smallest possible value

# Prediction Line Equation

- Computing *a* & *b* for the prediction equation
- To find the values of *a* & *b* that minimizes the sum of squared errors, use the following equations:

$$b = \frac{S_{yx}}{S_x^2}$$

$$a = \overline{Y} - b(\overline{X})$$

A: 1. Take the mean of Y & subtract it by the slope multiplied by the mean of X

B (slope): 1. Take covariance of X & Y, divide it by the variance of X

# How to Find Variance & Covariance

## Variance

1. Find the mean of X
2. Subtract each observation from the X category in the distribution from the X mean & square the value. Repeat for all values in the X section
3. Add up all those squared deviations
4. Divide that value by N-1 where N is the population size NOTE: This needs to be done for both X & Y

## Covariance

1. Find mean of X
2. Find mean of Y
3. For each case in the distribution, subtract that case's X by the mean of X and that case's Y by the mean of Y
4. Take those 2 numbers from step 3 and multiply them together
5. Repeat steps 3 & 4 for all the cases in the distribution then add them up
6. Divide by N-1 (where N is the population size

$$\text{The variance} = S^2_x = \frac{\Sigma (X-\overline{X})^2}{N-1}$$

$$\text{The Covariance } (X,Y) = S_{yx} = \frac{\Sigma (X-\overline{X})(Y-\overline{Y})}{N-1}$$

## Covariance

- The covariance is a measure of how X & Y vary together
- It basically tells us to what extent one variable goes together with the second variable
- The value reflects both strength & the direction of the relationship between the two variables
- Covariance can be positive or negative number. If the covariance is positive that means the two variables move together, when one increases so does the other
- If the covariance is negative then when one variable rises, the other variable goes down (inverse relationship)

## Example finding Covariance:

| Country | GNP per Capital (X) | % Willing to Pay (Y) |
|---|---|---|
| Denmark | 20.0 | 78% |
| Norway | 22.0 | 73% |
| Korea | 4.4 | 71% |

1. Find mean of X

Add up all the values in the X column, you get 46.4, then you divide 46.4 by 3 to get 15.4666 Rounding 15.4666 to the 1st decimal place, we get **15.5**

1. Find mean of Y

Add up all the values in the Y column, you get 221, then you divide 221 by 3 to get 73.666 Rounding 73.666 to the 1st decimal place, we get **74**

1. Subtract each X value in table from the mean of X

The equations we'd use for this would be:

- (20 - 15.5)
- (22 - 15.5)
- (4.4 - 15.5)

The values we'd get would be: 4.5, 6.5, -11.1

1. subtract each Y value in table from the mean of Y

The equations we'd use for this would be:

- (78-74)
- (73-74)
- (71-74)

The values we'd get would be: 4, -1, -3

1. Take those values then multiply the X values with its corresponding Y value, then add up the values after

The equations we'd use for this step would be:

- (4.5)(4)
- (6.5)(-1)
- (-11.1)(-3)

The values we'd get would be: 18, -6.5, 33.3 Now we add these values together & we get: 44.8

1. Now take that value of the top half of the equation and divide it by the population size (N) minus 1

Population size is 3 (Denmark, Norway, Korea) so our bottom half of th equation is 2 so it's 44.8/2 which gives us 22.4. **Our covariance of X & Y is 22.4**

# Example finding Variance:

| Country | GNP per Capital (X) | % Willing to Pay (Y) |
|---|---|---|
| Denmark | 20.0 | 78% |
| Norway | 22.0 | 73% |
| Korea | 4.4 | 71% |

1. We know the mean of X is 15.5 & the mean of Y is 74. If we didn't, you simply add up all the values in the corresponding category (if you were finding mean of X, add up all the values in X) then divide by the amount of numbers you added up (which in this case is 3).

2. Subtract the X value from the mean of X then square the value. Repeat this for all values in the X category

The equations we'd use for this in the X category are

*(20.0 - 15.5)^2* (22.0 - 15.5)^2 *(4.4 - 15.5)^2

The values we'd get for this are: 20.25, 42.25, 123.21

1. Add up all the values

Adding them all up, we get: 185.71

1. Divide by N-1, the population size - 1

Dividing by N (2), we get 92.855, so our **variance for X is 92.855, rounding up 2 decimal places, we get 92.86**

Repeating the 1st step for the Y values using the mean of Y:

The equations we'd use for this in the Y category are:

*(78 - 74)^2* (73 - 74)^2 *(71 - 74)^2

The values we'd get from this are: 16, 1, 9

Then we add up all the values to get: 26

Then we divide by 2 to get 13 so **our variance for Y is 13**

# Prediction equation example:

$$b = \frac{S_{yx}}{S_x^2}$$

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

B (Slope):

- Our bottom value is the variance of X (92.86)
- Our top value is the covariance of X & Y (22.4)
- 22.4/92.86 = **0.24**
- We rounded from the 2nd decimal place from 0.241

A (Y-intercept):

- The mean of Y is 74
- The mean of X is 15.5
- The b (slope) value is 0.24
- A = 74 - 0.24(15.5) so **A = 70.28**

# Equation for a straight line

- Using the values from prior calculations (above), we can plug in the values to the Y=a + b(X)
- **Y = 70.28 + 0.24(X)**

Using this chart again:

| Country | GNP per Capital (X) | % Willing to Pay (Y) |
|---|---|---|
| Denmark | 20.0 | 78% |
| Norway | 22.0 | 73% |
| Korea | 4.4 | 71% |

We can plugin the corresponding X values to our straight line equation can compute our **predicted values**:

Denmark: 75.08 Norway: 75.56 Korea: 71.34

- After finding the predicted values, we can plot these predicted values on the graph
- After plotting the predicted values, draw a line through them and this is your best-fitting line that minimizes errors throughout all observations in the scatter diagram.

# Summary: Properties of Regression Line

- Represents the predicted values for Y for any and all values of X
- Always goes through the point corresponding to the mean of both X & Y

- It is the best fitting line in that it minimizes the sum of the squared deviations (aka, sum of squared errors)
- Has a slope that can be positive or negative; null hypothesis is that the slope is zero

# The Coefficient of Determination or r^2

- r^2 measures the proportional reduction of error that results from using the linear regression model
- It represents the proportion of the total variation in the dependent variable Y, explained by the independent variable X
- r^2 ranges from 0.0 to 1.00
- tells us total variation of dependent variable that is explained by the independent variable
- 0 means independent variable doesn't explain variation in the dependent variable (no relationship)
- 1 means perfect deterministic relationship. Independent variable explains all variation in the dependent variable. You can perfectly predict the value of the dependent variable from the independent variable

# Calculating r^2

Equation:

$$r^2 = \frac{[\text{Covariance}(X,Y)]^2}{[\text{Variance}(X)][\text{Variance}(Y)]}$$

OR

$$r^2 = \frac{S^2_{xy}}{S^2_x S^2_y}$$

- take the covariance of X & Y then square it (numerator)
- multiply the variance of X & Y (denominator)

# The Correlation Coefficient (r)

- Pearson's Correlation Coefficient (r) - The square root of r^2. It's a measure of association between 2 interval-ratio variables
- Symmetrical measure - No specification of independent & dependent variables. You'll get the same Pearson's R.
- Ranges from -1.0 to +1.0. The sign (positive or negative) indicates direction. The closer the number is to positive or negative 1.0, the stronger the association between X & Y

- r = 0 means there is no association between the 2 variables
- r = +1 means a perfect positive correlation
- r = -1 means a perfect negative correlation
- You can just take the square root of the r^2 and boom. OR use the below equation:

$$r = \frac{\text{cov}(X,Y)}{[s.d.(X)][s.d.(Y)]}$$

- NOTE: to get the standard deviation, take the square root of the variance