Regression

Greg C Elvers

1

Correlation

- The purpose of correlation is to determine if two variables are linearly related to each other
- The correlation coefficient tells us:
  - the strength of the relation
  - the direction of the relation (direct or indirect)
- The correlation coefficient, however, does not tell us how the variables are related
  - I.e., it does not tell us how to predict the value of one variable given the value of the other

2

Regression

- The purpose of regression is to mathematically describe the relation between the variables
- Once you can describe the relation, you can predict the value of one variable given a value of the other variable
- When the variables are perfectly correlated, the prediction is perfect; the less correlated the variables, the less accurate the prediction

3

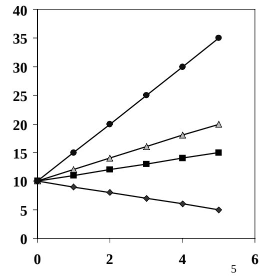
Regression Equation

- Because correlation assumes the variables are linearly related, the mathematical relation between the variables must be the equation of a line
- Y'=slope * X + intercept
- Y' (read Y prime) is the predicted value of the Y variable
- slope is how steep the line is
- intercept is where the line crosses the Y axis when X = 0

4

The Slope

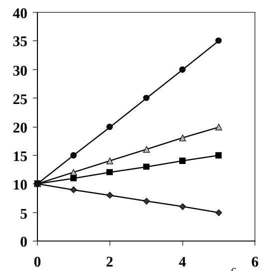
- The slope is how steep the line is
- The slope is defined as the change in the Y axis value divided by the change in the X axis value
- By just looking at the lines, which one has the steepest slope?



5

Slope

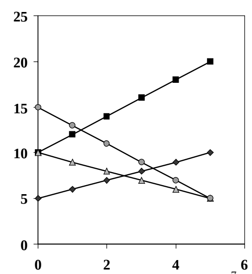
- Look at the left-most two points
  - For the blue line the change in Y is 15 - 10 = 5. The change in X is 1 - 0 = 1. The slope is 5 / 1 = 5
  - The slope of the green line is (12 - 10) / (1 - 0) = 2
  - Black's slope is 1
  - Red's slope is -1



6

Intercept

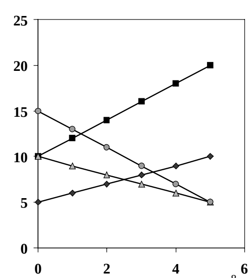
- The intercept is the Y axis value when X equals 0
  - It is where the line strikes the Y axis when X = 0
- Blue's intercept is 15
- Black and green's intercept is 10
- Red's intercept is 5



7

Equation of a Line

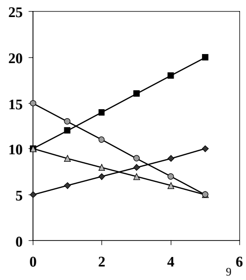
- To determine the equation of the black line, first determine its slope and intercept
- Slope = (12-10)/(1-0) = 2
- Intercept = 10
- Y' = 2 * X + 10



8

Equation of a Line

- Y’ = 2 * X + 10
- What value of Y is predicted when the value of X = 5?
- Y’ = 2 * 5 + 10 = 20
- Because the two variables are perfectly correlated, we can exactly predict the Y value given the X value



Regression When | r | < 1.0

- When the two variables are not perfectly correlated with each other, the points in a scatterplot will not fall directly on a line
- Thus, we will not be able to accurately predict the value of one variable given the value of the other variable
- The closer | r | is to 0, the less accurate our predictions will be

Determining Slope and Intercept when | r | < 1.0

- How do we determine the equation of the line when the data points do not fall on a line?
- We should try to find the line that does the best job of describing the data points
- That line is called the *line of best fit*, the *regression line*, or the *least squares line*; all three terms are synonymous

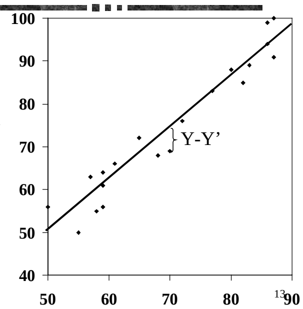
Line of Best Fit

- The line that we select as the regression line should minimize the errors that we make in our predictions
- The error in our prediction is given by:

$$\sum\left(Y - Y'\right)^2$$

$\Sigma(Y-Y')^2$

- What does this formula say?
- For each X, Y pair, calculate the predicted Y given X
- Subtract the predicted from the observed
- Square the difference
- Sum the squared differences



13

Why Square Y - Y'?

- You may wonder why we square the difference between the observed and predicted Y values
- The regression line (the line containing all the Y' values) is similar to the mean
- Recall that $\Sigma(X - \overline{X})^2$ was smaller than if we had substituted any other number for the mean
- That is, the mean minimizes the sum

14

Why Square Y - Y'?

- Thus, substituting Y' for the mean will make the squared errors smaller than if any other value was substituted

15

How To Determine the Slope

- The slope of the regression line should be influenced by three factors:
  - s_x
  - s_y
  - r

16

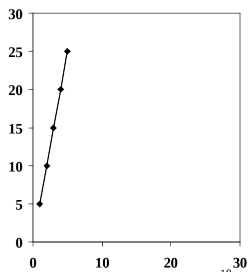
How To Determine the Slope

- The two standard deviations basically serve to standardize the difference in the variations of the two distributions
- The slope is proportional to the ratio: s_y / s_x
- The next several slides assume that X and Y are perfectly correlated

17

How To Determine the Slope

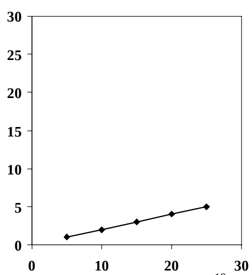
- If the standard deviation of X is small relative to the standard deviation of Y, then a small change in X should lead to a larger change in Y
- That is, the slope should be large (large ΔY / small ΔX)
- $s_y / s_x = 7.07 / 1.41 = 5$



18

How To Determine the Slope

- If the standard deviation of X is large compared to the standard deviation of Y, then a small change in X should lead to an even smaller change in Y
- The slope should be small (smaller ΔY / small ΔX)
- $s_y / s_x = 1.41 / 7.07 = 0.2$



19

How To Determine the Slope

- The slope also depends on the correlation of the two variables
- When the correlation is perfect, the slope is given by the ratio of the standard deviations
- When no correlation exists, the best prediction is always the mean no matter what the value of X is
- Thus, when r = 0, the slope should equal 0

20

How To Determine the Slope

- When | r | is between 0 and 1, the slope should be between 0 and s_y / s_x
- The closer r is to 0, the closer the slope should be to 0
- The closer | r | is to 1, the closer the slope should be s_y / s_x
- Thus, the slope is given by: slope = $r * s_y / s_x$

Computational Formula for Slope

- The computational formula for the slope of the regression line is:

$$\text{slope} = \frac{\sum XY - \left[\frac{(\sum X)(\sum Y)}{N}\right]}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

How To Determine the Intercept

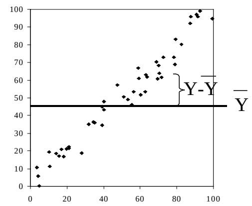
- Given that Y' = slope * X + intercept, $\overline{X}$, $\overline{Y}$, and r = 1, with a little algebra, we can solve for the intercept
- intercept = $\overline{Y}$ - slope * $\overline{X}$

Types of Variation in Regression

- There are three types of variation that are often mentioned when regression is discussed:
  - Total variation
  - Explained variation
  - Unexplained variation

Total Variation

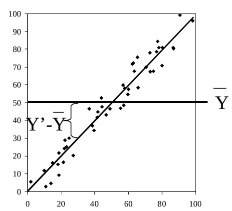
- The total variation is identical to the variation of the variable being predicted



$$s^2 = \frac{\sum\left(Y - \overline{Y}\right)^2}{N}$$

Explained Variation

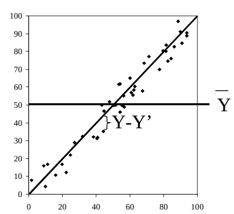
- The explained variation is the variation in Y that is can be explained by the regression equation



$$\text{Explained } s^2 = \frac{\Sigma\left(Y' - \overline{Y}\right)^2}{N}$$

Unexplained Variation

- The unexplained variation is the variation in Y that cannot be explained by the regression equation



$$\text{Unexplained } s^2 = \frac{\Sigma\left(Y - Y'\right)^2}{N}$$

Total Variation

- Total variation = explained variation + unexplained variation

$$\frac{\sum\left(Y - \overline{Y}\right)^2}{N} = \frac{\sum\left(Y' - \overline{Y}\right)^2}{N} + \frac{\sum\left(Y - Y'\right)^2}{N}$$

Partitioning of the Variance

- When we divide the total variance into two or more sub-totals, we are *partitioning the variance*
- This concept of dividing the total variation into different categories becomes an essential aspect of one of the most important inferential statistics, the ANalysis Of VAriance (ANOVA)

Coefficient of Determination

- The coefficient of determination, r^2, was defined as the proportion of variation in the Y data that was explainable by variation in the X data
- This can be given by the following formula

$$r^2 = \frac{\text{explained } s^2}{\text{total } s^2} = \frac{\frac{\sum\left(Y' - \overline{Y}\right)^2}{N}}{\frac{\sum\left(Y - \overline{Y}\right)^2}{N}}$$