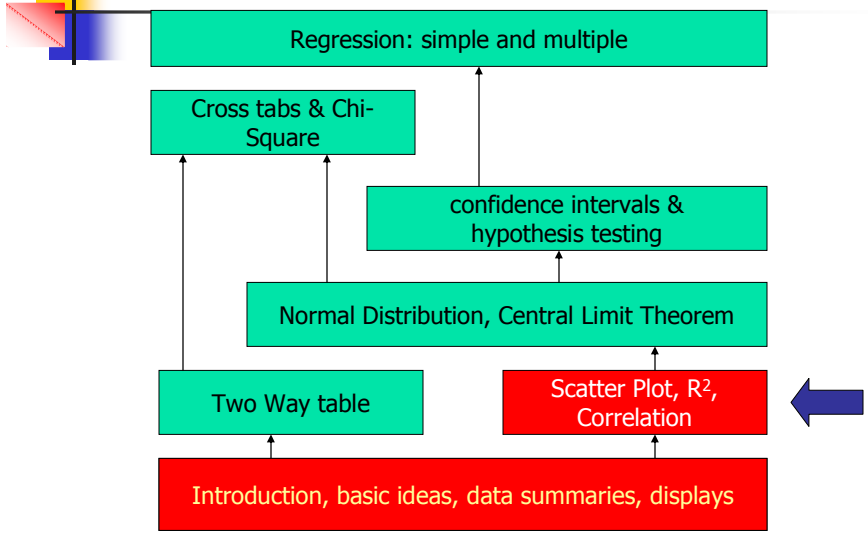
PUB – POS 316
Week 4

Scatterplots, regression & correlation

Navid Ghaffarzadegan
navidg@gmail.com
Last updated – Jan 1, 10

Course Road Map



Agenda

- Association
- Scatter plots
- Least Square Regression
- R-squared
- Correlation
  - (different from the book's structure in chapters 2-1, 2-2 and 2-3. But it does not mean that you don't need to read the book!)

Review from the last session

- Variance
- How far data points are from the average.
- What is the variance in the following data set: 2,4,3,2,3,4,3,3,3

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

Review from the last session

- Variance
- How far data points are from the average.
- What is the variance in the following data set: 2,4,3,2,3,4,3,3,3
- What is the standard deviation of the data set?

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

Review from the last session

- What is the standard deviation in:
  - 12, 12, 12, 12, 12, 12, 12, 12

Summary:

1. The variance is about how far the data points are from the average.
2. The standard deviation is the square root of variance.

Association

- We want to explain students' performance in PUB316.
- Central questions:
  - Examine grades.
  - Is there variation in grades?
  - What can explain the variation.
  - We love to explain variation in data.

Association

- Some variables can explain the variation in performance.
  - Doing homework
  - Hours spent to study
  - Class participation
  - ....
  - *We would like to find these variables (explanatory variables) to explain changes in the response variable (e.g., grade)*

Association

- Association between variables:

  Two variables are associated if knowing the value of one of them tells you something about the other one.

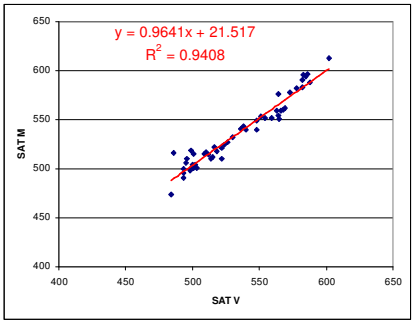
- Examples:
  - *Effort and grade*
    - *Positive association*
  - *Price and demand*
    - *Negative association*

Association – Scatter plot

- Example (2.6)
- We have data on SAT average scores in different states.
  - Q1: Is there any association between SATM and SATV?
    - We can look at the data.
    - We can draw a graph that helps us to see if there is an association. → scatter plot

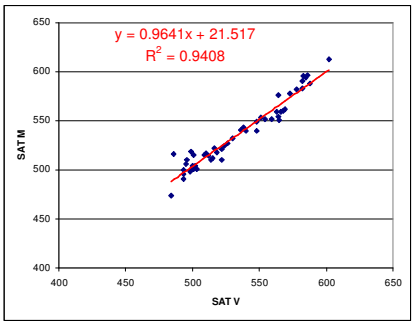
Association – Scatter plot

- Example (2.6): what does the graph say?



  - Q2: How can we predict SATM from SATV based on this data?
  - **If SATV=650 then SATM=?**

Association – Least Square Regression



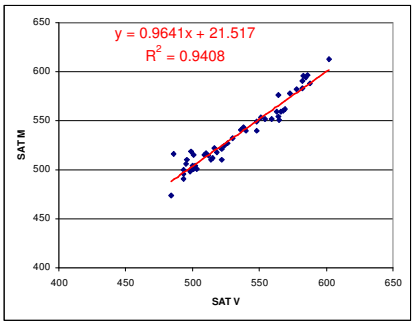
- Residual: $y_i - \hat{y}_i$
- The best fitting line minimizes the "sum of residual squared" $\sum (y_i - \hat{y}_i)^2$
- excel finds the line and we don't need to worry about it.

Association – Least Square Regression

Procedure 3: To find residuals

1. draw a scatter plot, 2. draw the trend line, 3. find the equation, 4. find y-hat (estimation from the equation), 5. find residuals

- Residual: $y_i - \hat{y}_i$
- Let's find residuals for this example.

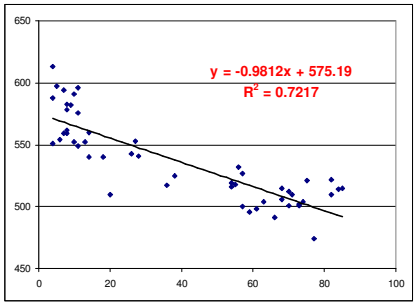


Association – Scatter plot

- Another Example
- We have data on SAT average scores in different states.
  - Q3: Is there any association between SATV and the percentage of people that take the test?
    - We can look at the data.
    - We can draw a graph that helps us to see if there is an association. → scatter plot

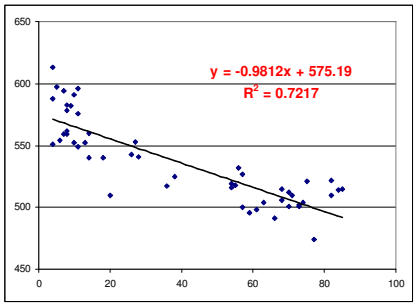
Association – Scatter plot

- Example: what does the graph say?



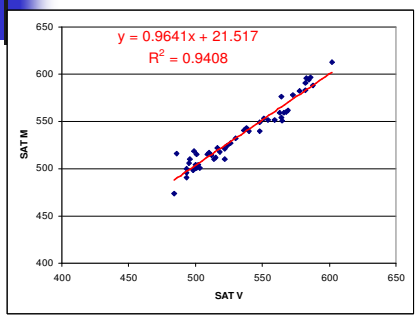
  - Q4: How can we predict SATV from Percentage of people who take the test?

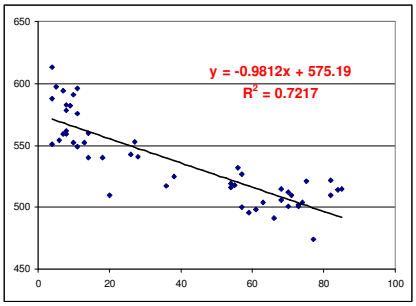
Association – Least Square Regression



- Residual: $y_i - \hat{y}_i$ Show me residuals!
- The best fitting line minimizes the "sum of residual squared" $\sum (y_i - \hat{y}_i)^2$
- excel finds the line and we don't need to worry about it.

Association – Least Square Regression





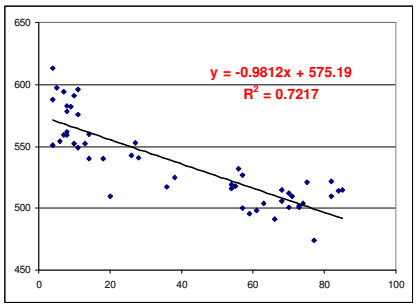
- Let's the graphs.. What do you see?
  - 1. Positive association vs. negative association
  - The left Graph shows a better fit.. Probably smaller residuals.
- We want to find a measure for fitness! Then we can systematically compare the graphs.

R- Squared

- We use regression lines to predict about the association between x and y.
  - For example: project size vs. needed budget
  - Health budget vs. health status of a state
  - Education vs. number of people in prison
  - etc.
- NOT all of these lines have a good explanatory power.
- And it is not easy to see how good a line fits the data.
- We need a measure for fitness.
- R-squared helps us.

R- Squared

- R^2 is the fraction of the variance of y's accounted for by the regression line.



- What is the variance of y? What is the variance of y-hat? What is the fraction?

R- Squared

- R^2 is the fraction of the variance of y's accounted for by the regression line.
- What is variance of y? What is variance of y-hat? What is the fraction?

$$R^2 = \frac{\frac{\sum(\hat{y}_i - \bar{y})^2}{n-1}}{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \ldots = \frac{b^2\sum(x_i - \bar{x})^2}{\sum(y_i - \bar{y})^2} = (b.\frac{S_x}{S_y})^2$$

R- Squared

- Question (from the spread sheet):
  - Which variable can explain SATM scores in a better way?
  - SATM (a response variable), what is good explanatory variable: GPA, HSM, HSS, SATV, SEX
  - 1. Draw scatter plots for GPA vs. SATM,
  - 2. Draw the trend line, report R-squared
  - 3. Continue these steps for some of the other variables vs. SATM
  - 4. Which one is better (which one has a higher R-squared?)
- Remember the definition of R-Squared.

R- Squared

- What is the maximum of R^2?
  - R^2 is always between 0 and 1.
- $R^2=1$ is the best fit. All points are on the regression line!
- Some exercise with scatter plots.
- The final point: remember: $R^2 = (b.\frac{S_x}{S_y})^2$
- We define correlation as: $r = b.\frac{S_x}{S_y}$

R- Squared

- R-Squared:
  - 1. Trend line reports
  - 2. RSQ(Array1, Array2)
- Correlation:
  - Correl(Array1, Array2)

Correlation

- How to calculate correlation?

Procedure 4: To calculate correlation from R-Squared

1. Take the square root of R^2.
2. The sign of correlation is the sign of the slope of the regression line.

- Find correlation between x and y in the following graph
- $R^2 = 0.72$
- → r= sqrt(0.72)
- Slope of the regression line is negative
- →r has a "negative" sign
- ➔ r= - 0.85