Regression & Prediction

Dr. Asheesh Srivastava
Professor, Head & Dean
Department of Educational Studies
School of Education,
Mahatma Gandhi Central University,
Motihari, East Champaran, Bihar-845401
profasheesh@mgcub.ac.in

If the variables in the bivariate distribution are related we shall find that the points in the scatter diagram will have tendency to cluster around some curve called the curve of regression.

If the curve is straight line, the curve is called the line of regression and there is said to be linear regression between variables.

Otherwise regression is curvilinear.

Regression

Definition

- **Regression Equation**

  Given a collection of paired data, the regression equation

$$(\hat{y} = a + bx)$$

algebraically describes the relationship between the two variables

- **Regression Line**

  (line of best fit or least-squares line)

  the graph of the regression equation

The Regression Equation

x **is the independent variable (predictor variable)**

$\hat{y}$ **is the dependent variable (response variable)**

$$\hat{y} = a + bx$$

Assumptions

1. **We are investigating only linear** relationships**.**

2. **For each x value, y is a random variable having a normal (bell-shaped) distribution. All of these y distributions have the same variance. Also, for a given value of x, the distribution of y-values has a mean that lies on the regression line. (Results are not seriously affected if we find departures from the assumptions of normal distributions and equal variances).**

Formula for y-intercept and slope

(slope)

Formula 1

$$b = \frac{n(\Sigma x_i y_i) - (\Sigma x_i)\ (\Sigma y_i)}{n(\Sigma x_i^2) - (\Sigma x_i)^2}$$

Formula 2

$$a = \bar{y} - b\ \bar{x}$$

(Intercept)

If you find r, then

Formula 3 slope = $b = r\, s_y/s_x$

where y is the mean of the $\bar{y}$-values and $\bar{x}$ is the mean of the x values

Formula 4 Intercept = $a = \bar{y} - b\bar{x}$

where $\bar{y}$ is the mean of the y-values, $\bar{x}$ is the mean of the x-values and b is the slope

The regression line fits the sample points best.

Residuals and the Least-Squares Property

Definitions

- **Residual**

  for a sample of paired (x,y) data, the difference ($y - \hat{y}$) between an observed sample y-value and the value of $\hat{y}$, which is the value of y that is predicted by using the regression equation.

- **Least-Squares Property**

  A straight line satisfies this property if the sum of the squares of the residuals is the smallest possible.

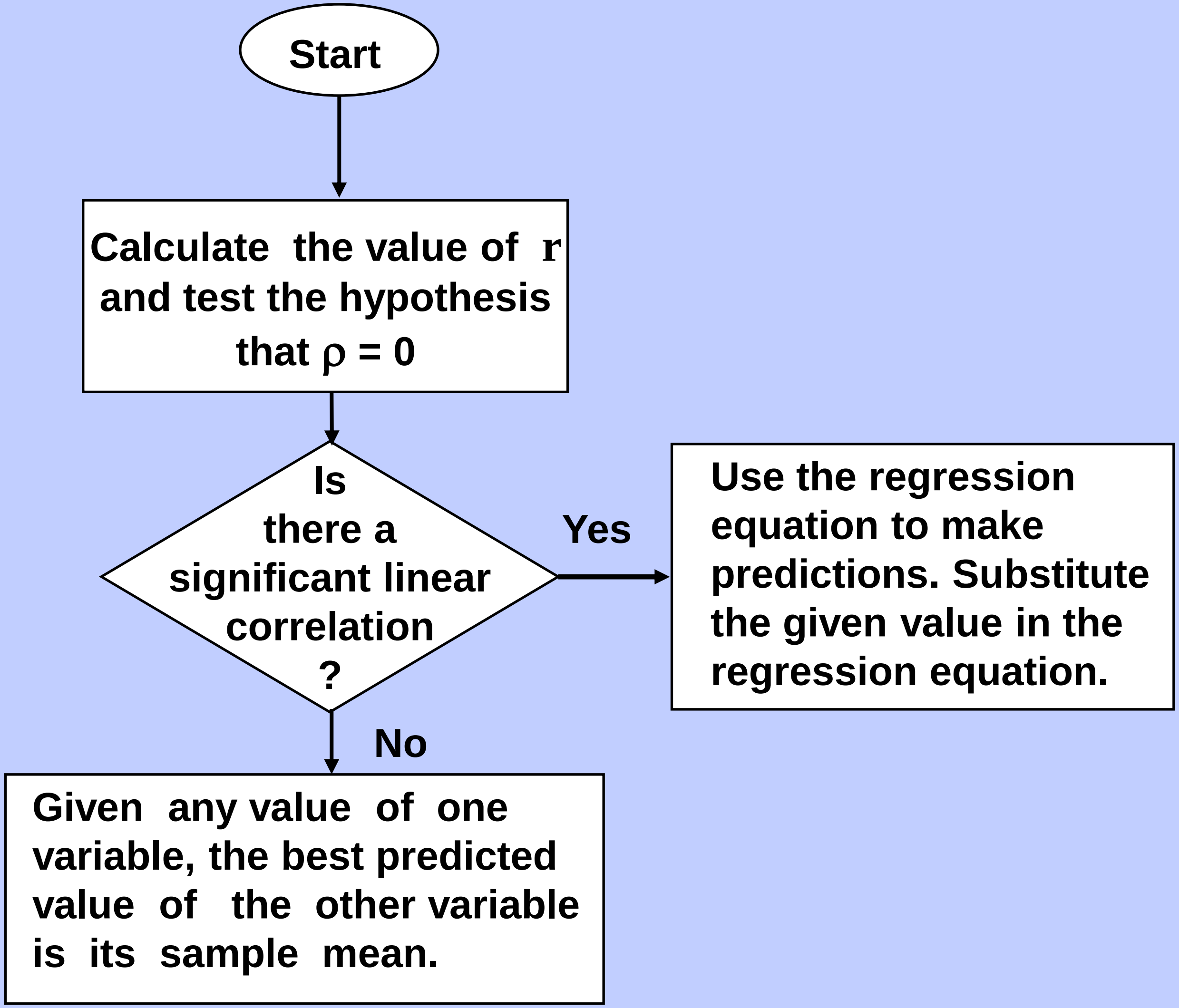
Predictions

In predicting a value of *y* based on some given value of *x* ...

1. **If there is not a significant linear correlation, the best predicted y-value is $\bar{y}$.**

2. **If there is a significant linear correlation, the best predicted y-value is found by substituting the x-value into the regression equation.**

Predicting the Value of a Variable



Guidelines for Using The Regression Equation

1. If there is no significant linear correlation, don’t use the regression equation to make predictions.

2. When using the regression equation for predictions, stay within the scope of the available sample data.

3. A regression equation based on old data is not necessarily valid now.

4. Don’t make predictions about a population that is different from the population from which the sample data was drawn.

Example

X	Y
1	34
2	36
3	37
4	39
5	41
10	50
15	59
18	64
20	68
30	86

Compute r, slope, intercept, regression

What is this equation used for?

What is the best predicted size of a household that discard 0.50 lb of plastic?

Data from the Garbage Project

x Plastic (lb)	0.27	1.41	2.19	2.83	2.19	1.81	0.85	3.05
y Household	2	3	3	6	4	2	1	5

Using a calculator:

$$a = 0.549$$

$$b = 1.48$$

$$y = 0.549 + 1.48\,(0.50)$$

$$y = 1.3$$

A household that discards 0.50 lb of plastic has approximately one person.

Definitions

- **Marginal Change**

  the amount a variable changes when the other variable changes by exactly one unit

- **Outlier**

  a point lying far away from the other data points

- **Influential Points**

  points which strongly affect the graph of the regression line

Multiple Regression

Definition

Multiple Regression Equation

A linear relationship between a dependent variable y and two or more independent variables ($x_1, x_2, x_3 \ldots, x_k$)

$$\hat{y} = m_0 + m_1x_1 + m_2x_2 + \ldots + m_kx_k$$

Generic Models

- **Linear:** $y = a + bx$
- **Quadratic:** $y = ax^2 + bx + c$
- **Logarithmic:** $y = a + b \ln x$
- **Exponential:** $y = ab^x$
- **Power:** $y = ax^b$
- **Logistic:** $y = \dfrac{c}{1 + ae^{-bx}}$

Development of a Good Mathematical Model

- **Look for a Pattern in the Graph:** Examine the graph of the plotted points and compare the basic pattern to the known generic graphs.
- **Find and Compare Values of R^2:** Select functions that result in larger values of R^2, because such larger values correspond to functions that better fit the observed points.
- **Think:** Use common sense. Don’t use a model that lead to predicted values known to be totally unrealistic.

Thank you