CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

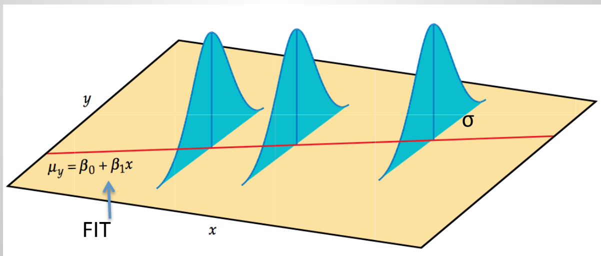
W6: 9/22-24/15

Administrative

- Updated Weekly Schedule (on Moodle)
- HW-2: Posted
  - 1st Due: 10/4/15 (Questions 1-7)
  - 2nd Due: 10/11/15 (Question 8, R-project)
- 1st Midterm: 10/6/15

Key Points From 9/22/15

- Population regression gives mean value. However, we can't observe this line on sample data. The statistical model for linear regression consists of population regression line and a description of the variation of *y* about the line. That is, data = fit + residual. The residual part represents deviations of the data from the line of population means. We assume that these deviations are Normally distributed with standard deviation σ.



Today

- Multiple Linear Regression
- Parameter Estimation

Regression Parameters

- Least-squares line:

$$\hat{y} = b_0 + b_1 x$$

- Intercept

$$b_0 = \overline{y} - b_1 \overline{x}$$

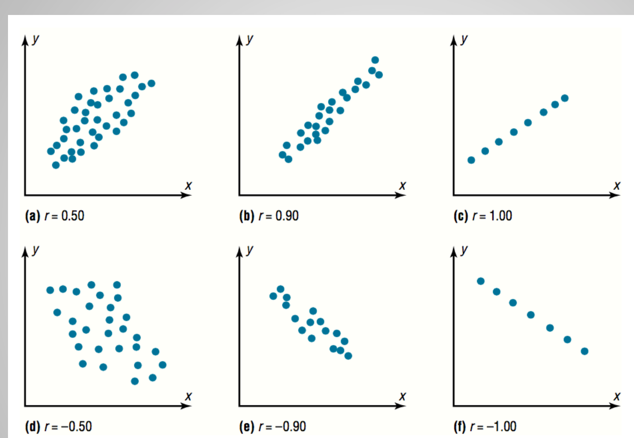
- Slope

$$b_1 = r \frac{s_y}{s_x}$$

- Residual, e_i = observed response – predicted response

$$e_i = y_i - b_0 - b_1 x_i$$

Correlation Coefficient is Important



$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

where n is the number of data pairs.

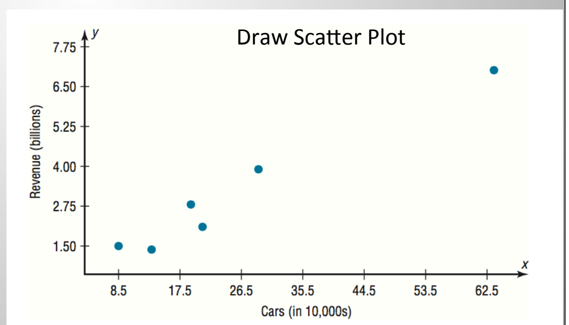
Various formulas, but this easy to compute

Ex-1: Car Rental Data

Company	Cars (in ten thousands)	Revenue (in billions)
A	63.0	$7.0
B	29.0	3.9
C	20.8	2.1
D	19.1	2.8
E	13.4	1.4
F	8.5	1.5

1st step to understand data

Draw Scatter Plot



Compute Correlation Coefficient

- Remember: *r* measures the strength and direction of a linear relationship between two variables.

NOTE: You need calculator for all exams, and only calculator is allowed

Step 1 Make a table as shown here.

Company	Cars x (in ten thousands)	Income y (in billions)	xy	x^2	y^2
A	63.0	7.0			
B	29.0	3.9			
C	20.8	2.1			
D	19.1	2.8			
E	13.4	1.4			
F	8.5	1.5			

Compute Correlation Coefficient

Step 2 Find the values of xy, x^2, and y^2 and place these values in the corresponding columns of the table.

The completed table is shown.

Company	Cars x (in 10,000s)	Income y (in billions)	xy	x^2	y^2
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	2.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

Step 3 Substitute in the formula and solve for r.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

$$= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982$$

The correlation coefficient suggests a strong relationship between the number of cars a rental agency has and its annual income.

Population (ρ) vs. Sample (r)

- The population correlation coefficient ρ is computed from taking all possible (*x*,*y*) pairs; it is designated by the Greek letter ρ (rho). The sample correlation coefficient (r) can then be used as an estimator of ρ if the following assumptions are valid.
  1. The variables *x* and *y* are *linearly* related.
  2. The variables are *random* variables.
  3. The two variables have a *bivariate normal distribution.*

NC STATE UNIVERSITY

The Significance of *r*

- The range of the correlation coefficient is between -1 and +1. When the value of *r* is near -1 or +1, there is a strong linear relationship. When the value of *r* is near 0, the linear relationship is weak or nonexistent.
- Since the value of *r* is computed from data obtained from samples, there are two possibilities when *r* is not equal to zero: either the value of *r* is high enough to conclude that there is a significant linear relationship between the variables, or the value of *r* is due to chance.

NC STATE UNIVERSITY

We can follow same 5-step process

- For rental car data, test if r is significant using α = 0.05. Use t test.

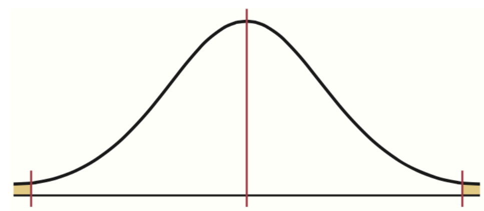
Step 1 State the hypotheses.

H_0: $\rho = 0$ and H_1: $\rho \neq 0$

NC STATE UNIVERSITY

We can follow same 5-step process

Step 2 Find the critical values. Since $\alpha = 0.05$ and there are $6 - 2 = 4$ degrees of freedom, the critical values obtained from Table F are ± 2.776, as shown



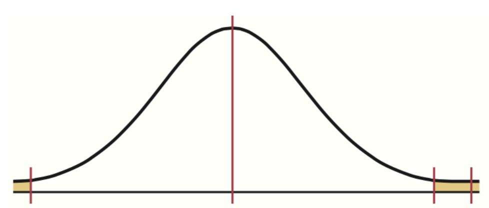
Step 3 Compute the test value.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = 0.982\sqrt{\frac{6-2}{1-(0.982)^2}} = 10.4$$

NC STATE UNIVERSITY

We can follow same 5-step process

Step 4 Make the decision. Reject the null hypothesis, since the test value falls in the critical region, as shown



Step 5 Summarize the results. There is a significant relationship between the number of cars a rental agency owns and its annual income.

For Car Rental Agency Data

- Compute Regression Line: y' = a + bx
- Recall formulae for a and b

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

where a is the y' intercept and b is the slope of the line.

9/28/15 © Raju Vatsavai CSC-591. 15

Computing Regression Line

The values needed for the equation are $n = 6$, $\Sigma x = 153.8$, $\Sigma y = 18.7$, $\Sigma xy = 682.77$, and $\Sigma x^2 = 5859.26$. Substituting in the formulas, you get

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{(18.7)(5859.26) - (153.8)(682.77)}{(6)(5859.26) - (153.8)^2} = 0.396$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{6(682.77) - (153.8)(18.7)}{(6)(5859.26) - (153.8)^2} = 0.106$$

Hence, the equation of the regression line $y' = a + bx$ is

$$y' = 0.396 + 0.106x$$

9/28/15 © Raju Vatsavai CSC-591. 16

NC STATE UNIVERSITY

Prediction Using Regression Line

- Let's say, x = 20.8, then what is the revenue
- y' = 2.6
- But actual y = 2.1 (Billions)
- That's significance difference, why?

NC STATE UNIVERSITY

Variation

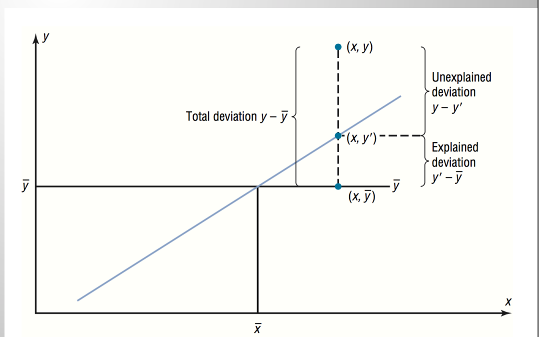
- Consider following simple data

x	1	2	3	4	5
y	10	8	12	16	20

- Then, y' = 4.8 + 2.8*x. (do this at home)*
- (x, y')
- 1, 7.6
- 2, 10.4
- 3, 13.2
- 4, 16.0
- 5, 18.8

Types of Variation

- The total variation: $\Sigma(y - \bar{y})^2$
  - Is the sum of the squares of the vertical distances each point is from the mean
- Has two components:
  - Explained variation (variation obtained from the relationship: $\Sigma(y' - \bar{y})^2$
  - Unexplained variation (variation due to chance): $\Sigma(y - y')^2$



Total Variation

- Total variation = Sum of explained variation + Sum of unexplained variation

$$\Sigma(y - \bar{y})^2 = \Sigma(y' - \bar{y})^2 + \Sigma(y - y')^2$$

- Using the simple data provided in previous slide, compute total variation
  - Answer = 92.8

NC STATE UNIVERSITY

Residual

- The values (y – y') are called residuals.
- A residual is the difference between the actual value of *y* and the predicted value *y′* for a given *x* value.
- The mean of the residuals is always 0.
- The sum of squares of the residuals computed using regression line is the smallest possible value.
- Therefore, a regression line is also called a least-squares line.

9/28/15 © Raju Vatsavai CSC-591. 21

NC STATE UNIVERSITY

Coefficient of Determination

- The coefficient of determination is the ratio of the explained variation to the total variation, denoted by r^2. Typically expressed as percentage.

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

- The coefficient of determination is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable. The symbol for the coefficient of determination is r^2.
- $(1-r^2)$ is called the coefficient of nondetermination

9/28/15 © Raju Vatsavai CSC-591. 22

Standard Error of the Estimate

- When a y' value is a predicted for a specific x value, the prediction is point prediction. However, we can construction a prediction interval about y' using the standard error of the estimate.
- The standard error of the estimate, denoted by s_{est}, is the standard deviation of the observed *y* values about the predicted *y'* values. The formula for the standard error of the estimate is

$$s_{\text{est}} = \sqrt{\frac{\Sigma(y - y')^2}{n - 2}}$$

Practice Example

- Based on the data collected (given below), secretary determines that there is significant relationship between age of copy machine and its monthly maintenance cost. Find the standard error of estimate.

Machine	Age x (years)	Monthly cost y
A	1	$ 62
B	2	78
C	3	70
D	4	90
E	4	93
F	6	103

Prediction Interval About y'

- From previous example, we can predict maintenance cost of 3-year old machine, but we don't know how accurate it is.
- Prediction interval is given by:

$$y' - t_{\alpha/2}s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x - \overline{X})^2}{n\,\Sigma x^2 - (\Sigma x)^2}} < y < y' + t_{\alpha/2}s_{\text{est}}\sqrt{1 + \frac{1}{n} + \frac{n(x - \overline{X})^2}{n\,\Sigma x^2 - (\Sigma x)^2}}$$

with d.f. $= n - 2$.

9/28/15 © Raju Vatsavai CSC-591. 25

Practice Example

- For the copy machine data, find the 95% prediction interval

9/28/15 © Raju Vatsavai CSC-591. 26

NC STATE UNIVERSITY

Multiple Regression

- In general, there will be more than one independent variable in the relationship.
- Multiple regression, explains the relationship between several independent variables and one dependent variable.

$$y' = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

NC STATE UNIVERSITY

Assumptions About Multiple Regression

The assumptions for multiple regression are similar to those for simple regression.

1. For any specific value of the independent variable, the values of the y variable are normally distributed. (This is called the *normality* assumption.)
2. The variances (or standard deviations) for the y variables are the same for each value of the independent variable. (This is called the *equal-variance* assumption.)
3. There is a linear relationship between the dependent variable and the independent variables. (This is called the *linearity* assumption.)
4. The independent variables are not correlated. (This is called the *nonmulticollinearity* assumption.)
5. The values for the y variables are independent. (This is called the *independence* assumption.)

NC STATE UNIVERSITY

Multiple Correlation Coefficient

- The strength of the relationship between the independent variables and the dependent variable is measured by a correlation coefficient, called multiple correlation coefficient, and is symbolized by *R*.

The formula for R is

$$R = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

where r_{yx_1} is the value of the correlation coefficient for variables y and x_1; r_{yx_2} is the value of the correlation coefficient for variables y and x_2; and $r_{x_1x_2}$ is the value of the correlation coefficient for variables x_1 and x_2.

NC STATE UNIVERSITY

Properties of R

- R ranges from 0 to +1.
  - Stronger relationship when R is close to +1
  - Weaker (or no) relationship when R is closer to 0.
- R is always higher than the individual correlation coefficients

NC STATE UNIVERSITY

Example: Board Exams

- Students GPA, Age, and State board score are given below.

Student	GPA x_1	Age x_2	State board score y
A	3.2	22	550
B	2.7	27	570
C	2.5	24	525
D	3.4	28	670
E	2.2	23	490

- The multiple regression equation is given by

$$y' = -44.81 + 87.64x_1 + 14.533x_2$$

9/28/15 CSC-591. 31

NC STATE UNIVERSITY

Example

- If GPA of a 25 year old student is 3.0, then what is the predicted state board score?

- Compute R

9/28/15 CSC-591. 32

NC STATE UNIVERSITY

Testing the Significance of R

- F test is used to test the significance of R.
- The hypothesis are:

$$H_0\colon \rho = 0 \qquad \text{and} \qquad H_1\colon \rho \neq 0$$

- F Test is given by:

The formula for the F test is

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

where n is the number of data groups $(x_1, x_2, \ldots, y)$ and k is the number of independent variables.

The degrees of freedom are d.f.N. $= n - k$ and d.f.D. $= n - k - 1$.

NC STATE UNIVERSITY

Example

- For student/state board data, test the significance at α=0.05

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}$$

$$= \frac{0.978/2}{(1 - 0.978)/(5 - 2 - 1)} = \frac{0.489}{0.011} = 44.45$$

The critical value obtained from Table with $\alpha = 0.05$, d.f.N. $= 3$, and d.f.D. $= 5 - 2 - 1 = 2$ is 19.16. Hence, the decision is to reject the null hypothesis and conclude that there is a significant relationship among the student's GPA, age, and score on the nursing state board examination.

NC STATE UNIVERSITY

Adjusted R^2

- Since R^2 is dependent on n (number of data pairs) and k (number of variables), often adjusted R^2 is used.

The formula for the adjusted R^2 is

$$R^2_{\text{adj}} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

NC STATE UNIVERSITY

Acknowledgements

- G. James, et. al., Moore, et. al.