ABILENE CHRISTIAN UNIVERSITY

Department of Mathematics

Linear Regression and Correlation

Section 15.5

Dr. John Ehrke
Department of Mathematics

Fall 2012

Introduction to Scatter Plots

When both variables to be displayed on a graph are quantitative, one variable can be plotted along the horizontal axis and the other along the vertical axis. The first variable is called x and the second called y, so that the graph takes the form of a plot of the points (x, y) corresponding to the two pieces of data for each measured variable. When each pair of data points is plotted the resulting figure is called a *scatter plot*.

- **What type of pattern do you see?** Is there a constant upward or downward trend? What shape does the trend follow? (i.e. linear, quadratic, etc..)
- **How strong is the pattern?** Do all of the points follow this pattern exactly or is the relationship weakly visible?
- **Are there any unusual observations?** Do the points cluster anywhere? Do we have any outliers?

Numerical Measures

Suppose we were hunting for a new house in a certain neighborhood and the realtor we have hired has provided us with a list of the last 12 residences sold in that area with their square footage and selling prices.

Residence	x (sq. ft)	y (thousands of dollars)
1	1360	\$178.5
2	1940	\$275.7
3	1750	\$239.5
4	1550	\$229.8
5	1790	\$195.6
6	1750	\$210.3
7	1600	\$205.2
8	1450	\$188.6
9	1870	\$265.7
10	2230	\$360.5
11	2210	\$325.3
12	1480	\$168.8

1. Create a scatter plot of this data.

2. From the scatter plot it appears the data forms a linear pattern. Using two points from the data set see if you can find a line which “fits” this trend well. What was your line?

3. Within each of the two sets of data, find the sample means $(\overline{x}, \overline{y})$ and standard deviations (s_x, s_y).

Linear Correlation Coefficient

In the previous slide we described each variable, x and y individually using descriptive statistics such as the mean and standard deviation. However, these measures do not describe the relationship between x and y for a particular residence–that is, how the size of the living space affects the selling price of the home. A simple measure that serves this purpose is called the *linear correlation coefficient*, denoted by r, and is defined as

$$r = \frac{s_{xy}}{s_x \cdot s_y}.$$

Notice that we have introduced a new quantity, s_{xy} called the *covariance*. The covariance is defined as

$$s_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

or equivalently,

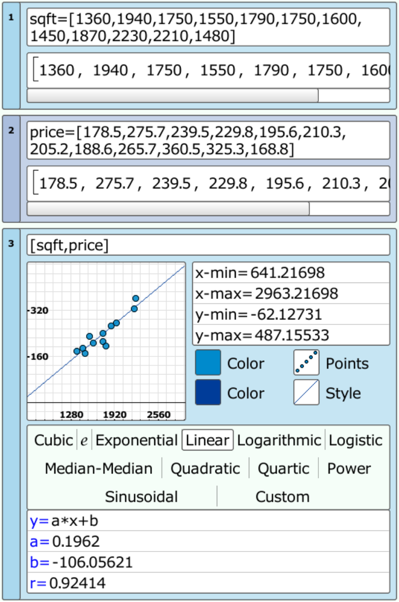
$$s_{xy} = \frac{\sum x_i y_i - \dfrac{\sum x_i \cdot \sum y_i}{n}}{n-1}.$$

This should be reminiscent of the computing formula for the standard deviation.

Correlation Between Square Footage and Home Price

Residence (i)	x_i (sq. ft)	y_i (thousands of dollars)	$x_i \cdot y_i$
1	1360	$178.5	
2	1940	$275.7	
3	1750	$239.5	
4	1550	$229.8	
5	1790	$195.6	
6	1750	$210.3	
7	1600	$205.2	
8	1450	$188.6	
9	1870	$265.7	
10	2230	$360.5	
11	2210	$325.3	
12	1480	$168.8	
Sum			

Calculating Linear Regressions via Calculator



1

sqft=[1360,1940,1750,1550,1790,1750,1600,
1450,1870,2230,2210,1480]

[1360 , 1940 , 1750 , 1550 , 1790 , 1750 , 160

2

price=[178.5,275.7,239.5,229.8,195.6,210.3,
205.2,188.6,265.7,360.5,325.3,168.8]

[178.5 , 275.7 , 239.5 , 229.8 , 195.6 , 210.3 , 2

3

[sqft,price]

x-min=641.21698
x-max=2963.21698
y-min=-62.12731
y-max=487.15533

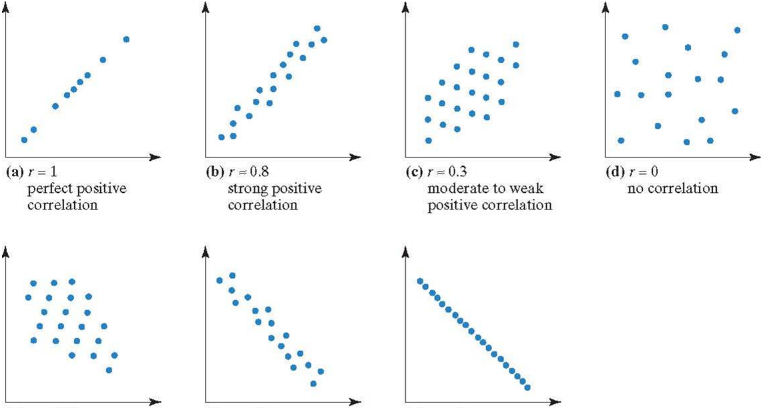
■ Color · Points
■ Color / Style

Cubic | e | Exponential | Linear | Logarithmic | Logistic
Median-Median | Quadratic | Quartic | Power
Sinusoidal | Custom

y=a*x+b
a=0.1962
b=-106.05621
r=0.92414

Possible Values of r

Based on our observations from the previous example, what values can r have?



(a) $r = 1$
perfect positive correlation

(b) $r \approx 0.8$
strong positive correlation

(c) $r \approx 0.3$
moderate to weak positive correlation

(d) $r = 0$
no correlation

(e) $r \approx -0.3$
moderate to weak negative correlation

(f) $r \approx -0.8$
strong negative correlation

(g) $r = -1$
perfect negative correlation

The value of r^2

Example

Ten different second-year medical students took blood pressure measurements of the same patient and the results are listed below.

Systolic	138	130	135	140	120	125	120	130	130	144
Diastolic	82	91	100	100	80	90	80	80	80	98

Calculate r and r^2 for this data and interpret your results.

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

Linear Regression Equations

If we suspect a correlation between a set of data points it is natural to wonder whether or not we can use the suspected relationship to predict values of say y given a particular x. The answer is yes, but our estimation is only as reliable as the correlation coefficient predicts it to be. The equation of the straight line that best represents the relationship between x and y is called the regression line, and its equation is called the regression equation. This also sometimes called the *line of best fit*, or the *least squares line*.

Example

Consider the set of blood pressure readings from the previous example. Use this data to find and plot the line of best fit. Predict the diastolic pressure reading for a person with a systolic reading of 110.

Systolic	138	130	135	140	120	125	120	130	130	144
Diastolic	82	91	100	100	80	90	80	80	80	98

Polling Question #23

Listed below are the weights (in pounds) and the highway fuel consumption amounts in miles per gallon of 7 randomly selected automobiles.

Weight (lbs)	3175	3450	3225	3985	2440	2500	2290
Fuel Consumption (mpg)	17	15	20	14	27	30	37

What is the linear correlation coefficient r, and linear regression equation for this set of data?

(a) $r = 0.92$, $y = 0.01287x + 61.57177$

(b) $r = -0.92$, $y = -0.01287x + 61.57177$

(c) $r = -0.92$, $y = -66.231x + 4523.14$

(d) $r = -0.5$, $y = 17x + 3175$

Polling Question #24

Listed below are the weights (in pounds) and the highway fuel consumption amounts in miles per gallon of 7 randomly selected automobiles.

Weight (lbs)	3175	3450	3225	3985	2440	2500	2290
Fuel Consumption (mpg)	17	15	20	14	27	30	37

If the linear regression model for this set of data is $y = -0.01287x + 61.57177$ what fuel consumption in MPG is predicted for a car with a weight of 5000 lbs? Does this answer make sense?

(a) -2.7 mpg

(b) -383716 mpg

(c) 10 mpg

(d) 2.7 mpg