Correlation, Regression & Error

Dr. Bashar Al-Shboul
The University of Jordan

- The correlation coefficient, *r,* measures the strength and direction of the linear relationship between two quantitative variables.
  - The formula to find the correlation coefficient is

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

- The point *(x', y')* is the *(mean of x, mean of y)*
  - The values s_x and s_y are the individual standard deviations of *x* and *y* respectively.
  - *n* represents the number of data pieces.

- Facts about Correlation:
  - Positive *r* indicates positive association and negative *r* indicates negative association between variables.
  - *r* is always between -1 and 1.
  - The closer I *r* I is to 1, the stronger the association. A weak association will have an *r* value close to 0.
  - Correlation is strongly influenced by outliers

Monopoly - Correlation

Property	Spaces from GO	Cost
Mediterranean Avenue	1	60
Baltic Avenue	3	60
Reading Railroad	5	200
Oriental Avenue	6	100
Vermont Avenue	8	100
Connecticut Avenue	9	120
St. Charles Place	11	140
Electric Company	12	150
States Avenue	13	140
Virginia Avenue	14	160
Penn Railroad	15	200

LINEAR REGRESSION

- Regression line is a line that describes the relationship between the explanatory variable *x* and the response variable *y.*
  - The least squares regression line formula is $y = b.x + a$ where *a* is the y-intercept and *b* is the slope.
  - The slope of a regression line can help determine if a relationship exists between two variables. When the slope of a regression line is zero, no relationship exists

- Regression lines can be used to predict a value for *y* given a value of *x.*
- The least squares regression line (or LSRL) is a mathematical model used to represent data that has a linear relationship.
- The slope, *b,* is calculated with the formula $b=r\ (Sy/Sx)$
- And the y-intercept is $a = y' - b.x'$

- Notice that the formula for slope is

$$b = r\left(\frac{s_y}{s_x}\right)$$

- This means that a change in one standard deviation in *x* corresponds to a change of *r* standard deviations in *y.*
- In other words, we can say that on average, for each unit increase in *x,* then is an increase (or decrease if slope is negative) of I *b* I units in *y.*

Monopoly - Correlation

Property	Spaces from GO	Cost
Mediterranean Avenue	1	60
Baltic Avenue	3	60
Reading Railroad	5	200
Oriental Avenue	6	100
Vermont Avenue	8	100
Connecticut Avenue	9	120
St. Charles Place	11	140
Electric Company	12	150
States Avenue	13	140
Virginia Avenue	14	160
Penn Railroad	15	200

- The LSRL can be used to predict values of *y* given values of *x.*
  - We need to be careful when predicting. when we are estimating *y* based on values of *x* that are much larger or much smaller than the rest of the data, this is called extrapolation.
- Use the LSRL found in previous example to predict the cost of a property that is 50 spaces from GO

- The square of the correlation *(r),* is called the coefficient of determination.
  - It is the fraction of the variation in the values of *y* that is explained by the regression line and the explanatory variable.
  - When asked to interpret r^2 we say, "approximately $r^2(100)$% of the variation in *y* is explained by the LSRL of *y* on *x* ."

- Facts about the coefficient of determination:
  - The coefficient of determination is obtained by squaring the value of the correlation coefficient.
  - The symbol used is r^2
  - Note that $0 < r^2 < 1$
  - r^2 values close to 1 would imply that the model is explaining most of the variation in the dependent variable and may be a very useful model.
  - r^2 values close to 0 would imply that the model is explaining little of the variation in the dependent variable and may not be a useful model.

- The following 9 observations compare the *x* (a measure of body build) and dietary energy density, *y.*

X	221	228	223	211	231	215	224	233	268
Y	.67	.86	.78	.54	.91	.44	.9	.94	.93

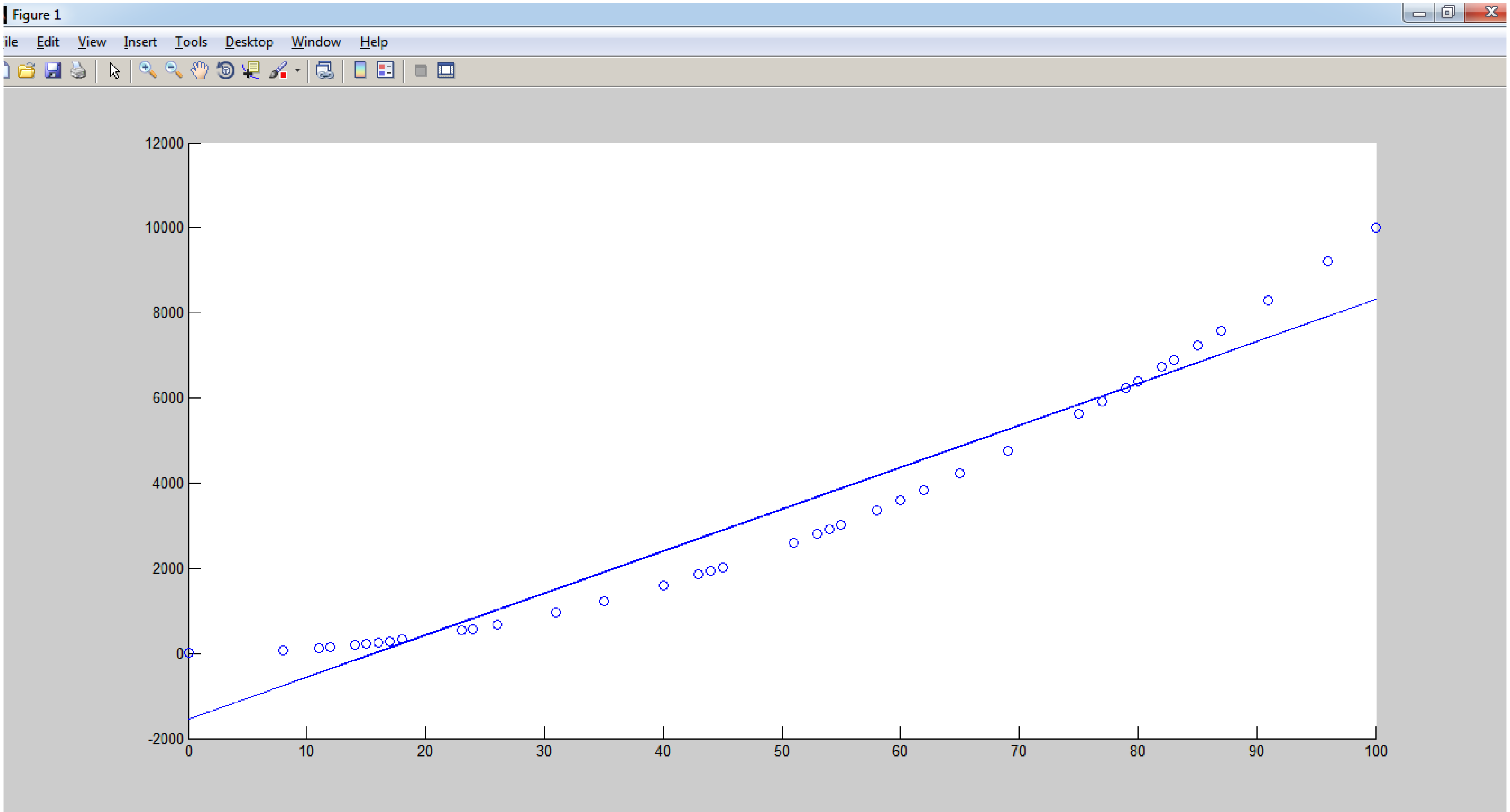
- Make a scatter-plot of the data
- *Compute Regression Line*
- Provide an interpretation of the slope of this line in the context of these data
- Find the correlation coefficient for the relationship. Interpret this number
- Find the coefficient of determination for the relationship. Interpret this number

Example - pricing a house

Table 1. House values for regression model

House size (square feet)	Lot size	Bedrooms	Granite	Upgraded bathroom?	Selling price
3529	9191	6	0	0	$205,000
3247	10061	5	1	1	$224,900
4032	10150	5	0	1	$197,900
2397	14156	4	1	0	$189,900
2200	9600	4	0	1`	$195,000
3536	19994	6	1	1	$325,000
2983	9365	5	0	1	$230,000
3198	9669	5	1	1	????

Example – Regression Line



In MATLAB – Sample Fit

- a = rand(50,2);
- a(:,1) = uint16(a(:,1) * 100);
- a(:,2) = a(:,1).^2 + 4;
- scatter(a(:,1), a(:,2));
- aa = dataset(a(:,1), a(:,2), 'Varnames',{'X','Y'});
- y = fitlm(aa);
- hold on, plot(a(:,1), y.Fitted);

ERRORS & ACCURACY MEASURES

Predictor Error Measures

- Measure predictor accuracy: measure how far off the predicted value is from the actual known value
- **Loss function**: measures the error between y_i and the predicted value y_i'
  - Absolute error: $| y_i – y_i' |$
  - Squared error: $(y_i – y_i')^2$

Four different types of error measures can be used, as follows:

Mean absolute error: $$\frac{\sum_{i=1}^{d} |y_i - y_i'|}{d}$$

Relative absolute error: $$\frac{\sum_{i=1}^{d} |y_i - y_i'|}{\sum_{i=1}^{d} |y_i - \bar{y}|}$$