Linear Regression

Ways to obtain a best fit line

- In a calculator, put x in L1 and y in L2. Choose STAT/CALC/LIN REG L1, L2, (optional) Y1 (VARS/Y-Vars/1/1).
- From computer output, find the COEF column. The y-intercept is the coefficient labeled CONSTANT, and the slope is the coefficient of the explanatory variable.
- Use the formula $b_1 = r\frac{s_y}{s_x}$ to find the slope and $b_0 = \bar{y} - b_1\bar{x}$ to get the y-intercept.

Properties of the correlation coefficient, r

- r tells the strength and direction of a *linear* relationship.
- r can only be calculated for graphs with 2 numerical (quantitative) variables.
- r is always between -1 and 1, inclusive.
- Graphs with positive slopes have positive r values; graphs with negative slopes have negative r values.
- r remains unchanged if x and/or y are rescaled.
- r remains unchanged if x and y are interchanged.
- r is dimensionless (has no units).
- r is not resistant to the effects of outliers.

Residuals

- To find a residual, subtract the predicted y-value from the actual y-value
  $$\text{residual} = y - \hat{y}$$
- The mean of the residuals is 0.
- The best fit, or least squares, line minimizes the sum of the squares of the residuals.
- A residual plot shows the residuals on the y-axis and the explanatory variable or the predicted y-values on the x-axis.
- Points with large residuals are called outliers. Points which change the slope of the line and the correlation coefficient greatly when removed are called influential points.

Is a relationship linear?

- Start with a scatterplot of the data points. Does it look linear?
- Examine the residual plot, if available. If it does not have a pattern, then x and y have a linear relationship.
- Do a linear regression t test. (2^{nd} Semester)

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.

How to interpret values in context

- Slope: For every (increase, decrease) of one (unit) in (context of x), there is an average (increase, decrease) in (context of y) of (slope)(units).

  Example: y = height of a plant in cm, x = age in months, where $\hat{y} = 1.2 + 2.3x$
  For every additional month, there is an average increase in the plant's height of 2.3 cm.

- Y-intercept: When the (context of x) is 0 (unit), I would predict that the (context of y) would be (y-intercept).

  Example: y = height of a plant in cm , x = age in months, where $\hat{y} = 1.2 + 2.3x$
  When the plant is 0 months old, I would predict that the height would be 1.2 cm.
  (Remember the y-intercept may not be a meaningful value, like this one)

- Correlation coefficient (r): The correlation coefficient of _____ indicates that there is a (strong, moderate, weak), (positive, negative) linear relationship between (context of y) and (context of x).

  Example: height of plant $r = 0.945$
  The correlation coefficient of 0.945 indicates that there is a strong positive linear relationship between the age of the plant and its height.

- Coefficient of determination (r^2): (r^2) % of the variability in (context of y) can be explained by the linear association with (context of x)

  Example: height of plant $r = 0.945$ $r^2 = .893$
  89.3% of the variability in the height of the plant can be explained by the linear association with the age of the plant.

- Residual plot: The residual plot (is randomly scattered, has a pattern) indicating that a linear model (is, is not) appropriate.

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.

Multiple Choice Questions on Linear Regression

1. Residuals are

(A) possible models not explored by the researcher.
(B) variation in the response variable that is explained by the model.
(C) the difference between the observed response and the values predicted by the model.
(D) data collected from individuals that is not consistent with the rest of the group.
(E) a measure of the strength of the linear relationship between x and y

2. Data was collected on two variables x and y and a least squares regression line was fitted to the data. The resulting equation is $\hat{y} = -2.29 + 1.70x$. What is the residual for point (5, 6)?

(A) −2.91
(B) −0.21
(C) 0.21
(D) 6.21
(E) 7.91

3. Child development researchers studying growth patterns of children collect data on the heights of fathers and sons. The correlation between the fathers' heights and the heights of their 16-year-old sons is most likely to be…

(A) near −1.0
(B) near 0
(C) near +0.7
(D) exactly +1.0
(E) somewhat greater than +1.0

4. Given a set of ordered pairs (x, y) with $s_x = 2.5, s_y = 1.9$, $r = .63$, what is the slope of the regression line of y on x?

(A) 0.48
(B) 0.65
(C) 1.32
(D) 1.90
(E) 2.63

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.

5. The relation between the selling price of a car (in \$1,000) and its age (in years) is estimated from a random sample of cars of a specific model. The relation is given by the following formula:

$$\text{SellingPrice} = 24.2 - (1.182)\text{Age}$$

Which of the following can be concluded from this equation?

(A) For every year the car gets older, the selling price drops by approximately \$2420.
(B) For every year the car gets older, the selling price goes down by approximately 11.82 percent.
(C) On average, a new car costs about \$11,820.
(D) On average, a new car costs about \$23,018.
(E) For every year the car gets older, the selling price drops by approximately \$1182.

6. All but one of these statements is false. Which one could be **true**?

(A) The correlation between a football player's weight and the position he plays is 0.54.
(B) The correlation between a car's length and its fuel efficiency is 0.71 miles per gallon.
(C) There is a high correlation (1.09) between height of a corn stalk and its age in weeks.
(D) The correlation between the amounts of fertilizer used and quantity of beans harvested is 0.42.
(E) There is a correlation of 0.63 between gender and political party.

7. It is easy to measure the circumference of a tree's trunk, but not so easy to measure its height. Foresters developed a model for ponderosa pines that they use to predict tree's height (in feet) from the circumference of its trunk (in inches):

$$\ln \hat{h} = -1.2 + 1.4(\ln C)$$

A lumberjack finds a tree with a circumference of 60 inches, how tall does this model estimate the tree to be?

(A) 5 ft
(B) 11 ft
(C) 19 ft
(D) 83 ft
(E) 93 ft

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.

8. Which is true?

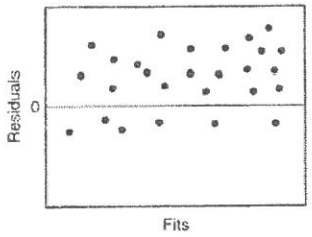
I. Random scatter in the residuals indicates a linear model.
II. If two variables are very strongly associated, then the correlation between them will be near +1.0 or −1.0.
III. Changing the units of measurement for x or y changes the correlation coefficient.

(A) I only
(B) II only
(C) I and II only
(D) II and III only
(E) I, II, and III

9. If the coefficient of determination r^2 is calculated as 0.49, then the correlation coefficient

(A) cannot be determined without the data
(B) is − 0.70
(C) is 0.2401
(D) is 0.70
(E) is 0.7599

10. Which of the following is a correct conclusion based on the residual plot displayed?



(A) The line overestimates the data.
(B) The line underestimates the data.
(C) It is not appropriate to fit a line to these data since there is clearly no correlation.
(D) The data are not related.
(E) There is a nonlinear relationship between the variables.

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.

Free Response Questions on Linear Regression

1. The National Directory of Magazines tracks the number of magazines published in the United States each year. An analysis of data from 1988 to 2007 gives the following computer output. The dates were recorded as years since 1988. Thus, the year 1988 was recorded as year 0. A residual plot (not shown) showed no pattern.

Predictor	Coef	StDev	T	P
Constant	13549.9	2.731	7.79	0.000
Years	325.39	0.1950	10.0	0.000
S = 836.2	R-Sq = 84.8%	R-Sq (adj) = 80.6%		

(a) What is the value of the slope of the least squares regression line? Interpret the slope in the context of this situation.

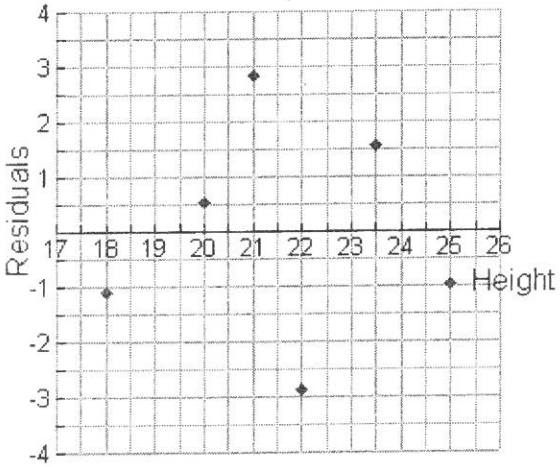
(b) What is the value of the y-intercept of the least squares regression line? Interpret the y-intercept in the context of this situation.

(c) Predict the number of magazines published in the United States in 1999.

(d) What is the value of the correlation coefficient for number of magazines published in the US and years since 1988? Interpret this correlation.

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.

2. The heights (in inches) and weights (in pounds) of six male Labrador Retrievers were measured. The height of a dog is measured at the shoulder. A linear regression analysis was done, and the residual plot and computer output are given below.



Predictor	Coef	StDev	T	P
Constant	−13.430	1.724	7.792	0.0000
Height	3.6956	0.4112	8.987	0.0004
S = 2.297	R-Sq = 95.3%	R-Sq (adj) = 90.6%		

(a) Is a line an appropriate model to use for these data? What information tells you this?

(b) Write the equation of the least squares regression line. Identify any variables used in this equation.

(c) Dakota, a male Labrador, was one of the dogs measured for this study. His height is 23.5 inches. Find Dakota's predicted weight **and** Dakota's actual weight.

Copyright © 2008 Laying the Foundation, Inc., Dallas, Texas. All rights reserved.
These materials may be used for face-to-face teaching with students only.