Unit 02 Review

Ways to obtain a line of best fit

IF YOU HAVE DATA VALUES:

1. In your calculator, choose STAT > 1.EDIT and enter your x values into L1 and your y values into L2
2. Choose STAT > CALC > 8. Linreg(a + bx) and choose L1 and L2
3. Your calculator will give you values for a and b. If you have turned DiagnosticOn (using the Catalog feature), you will also be given values for r and R^2.

IF YOU HAVE A MINITAB OUTPUT:

1. Look at the Coef column.
    a. The CONSTANT row tells you the y-intercept (a)
    b. The VARIABLE row tells you the slope (b)
2. Look for R-Sq to get the R^2 value. Do NOT use R-Sq (adj).

IF YOU HAVE SUMMARY STATISTICS:

1. Use the following formulas (also on your equation sheet):

slope (b): $b_1 = r\frac{s_y}{s_x}$

y-intercept (a): $b_0 = \bar{y} - b_1\bar{x}$

Properties of the correlation coefficient, r

- r tells us the strength and direction of a *linear* relationship between x and y
- 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

Is a relationship linear?

- Start with a scatterplot of 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 (Unit 08)

Residuals

- To find a residual, subtract the predicted y-value from the actual y-value:
  - 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 variables or the predicted y-values on the x-axis.
- Points with large residuals are called outliers (in the y-direction). Points which change the slope of the line and the correlation coefficient greatly when removed are called influential points.
- To plot residuals on your calculator when given lists of data, run linear regression (STAT>CALC>8. Linreg(a + bx)) and then use your STAT PLOT to plot L1 vs. Resid (can be found by clicking 2nd LIST).

How to interpret values in context

- Slope: For every [increase/decrease] in one [unit] of [context of x], there a predicted [increase/decrease] of [context of y] of [slope] [units].

- Y-intercept: When the [context of x] is 0 [unit], the predicted [context of y] would be [y-intercept].

- 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].

- Coefficient of determination (R^2): ___% of the variation in the values of [y in context] can be explained/accounted for by the LSRL of [context of y] on [context of x].

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

Transforming to get a linear model

- When a graph of y vs. x does not appear liner, either y or x or both may be transformed (for example, by taking the log or raising to a power) in order to get a linear graph.
- When using transformed models to make predictions, substitute x into the equation and perform the inverse operation/transformation to get the predicted y value.
- Common transformations: Exponential (x vs. logy) and Power (logx vs. logy)

Multiple Choice:

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

C

2. Data was collected on two variables x and y and a least squares regression line was fitted to the data. The resulting equation $\hat{y} = -2.29 + 1.70x$. What is the residual for the point (5,6)?
   a. -2.91
   (b.) -0.21
   c. 0.21
   d. 6.21
   e. 7.91

B

$\hat{y} = -2.29 + 1.70(5) = 6.21$

$\text{residual} = y - \hat{y} = 6 - 6.21 = -0.21$

3. Child development researchers studying growth patterns of children collect data on the heights of fathers and sons. The correlation between fathers' heights and the heights of their 16-year-old sons is mostly likely to be...
   ~~a.~~ Near -1.0
   ~~b.~~ Near 0
   (c.) Near +0.7
   ~~d.~~ Exactly +1.0
   ~~e.~~ Somewhat greater than +1.0

C

4. Given a set of ordered pairs (x, y) with $s_x = 2.5$, $s_y = 1.9$, $r = 0.63$, 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

A

$b = r\frac{s_y}{s_x} = 0.63\frac{1.9}{2.5} = 0.4788$

5. The relation between the selling price of a car (in thousands of dollars) 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:

$$\widehat{selling\ price} = 24.2 - 1.182(\text{age})$$

Which of the following can be concluded from this equation?

E

a. For every year the car gets older, the selling price goes down by approximately $2420
b. For every year the car gets older, the selling price goes down by approximately 11.82%
c. On average, a new car costs about $11820
d. On average, a new car costs about $23018
(e.) For every year that 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?

D

a. The correlation between a football player's weight and the position he plays is 0.54 not both Quantitative
b. The correlation between a car's length and its fuel efficiency is 0.71 miles per gallon r has no units
c. There is a high correlation (1.09) between height of a corn stalk and its age in weeks $-1 < r < 1$
(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):

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

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

E

a. 5 ft
b. 11 ft
c. 19 ft
d. 83 ft
(e.) 93 ft

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

$\hat{h} = e^{-1.2} \cdot C^{1.4} = e^{-1.2} \cdot 60^{1.4} = 92.95 ft$

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 +1.0 or -1.0 ✓
✗ III. Changing the units of measurement for x or y changes the correlation coefficient ✗

C

a. I only
b. II only
(c.) I and II only
d. II and III only
e. I, II, and III

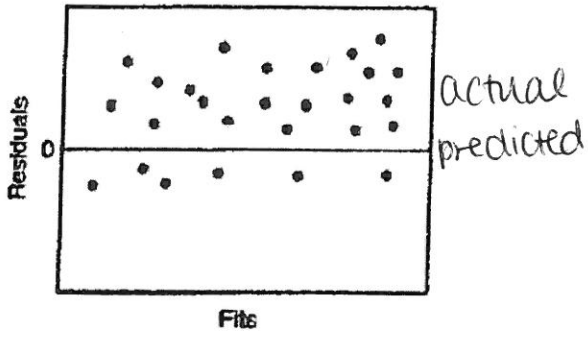
9. If the coefficient of determination r^2 is calculated as 0.49, then the correlation coefficient

(a.) Cannot be determined without the data
b. Is -0.70
c. Is 0.2401
d. Is 0.70
e. Is 0.7599

A

we don't know if r is positive or negative

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



B

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

Free Response:

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 2008 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.

The slope is 325.39 magazines per year.

For each year since 1988, the predicted number of magazines published in the US increases by about 325.

(you can substitute "predicted" with "on average")

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.

The y-intercept is 13549.9 magazines.

The predicted number of magazines published in the US in 1988 (year 0) is 13550 magazines.

c. Predict the number of magazines published in 1999. year 11 = 1999-1988

$\widehat{\text{magazines}} = 13549.9 + 325.39(\text{years since 1988})$

$\widehat{\text{magazines}} = 13549.9 + 325.39(11) = 17129$

We predict that there were 17129 magazines published in the US in 1999.

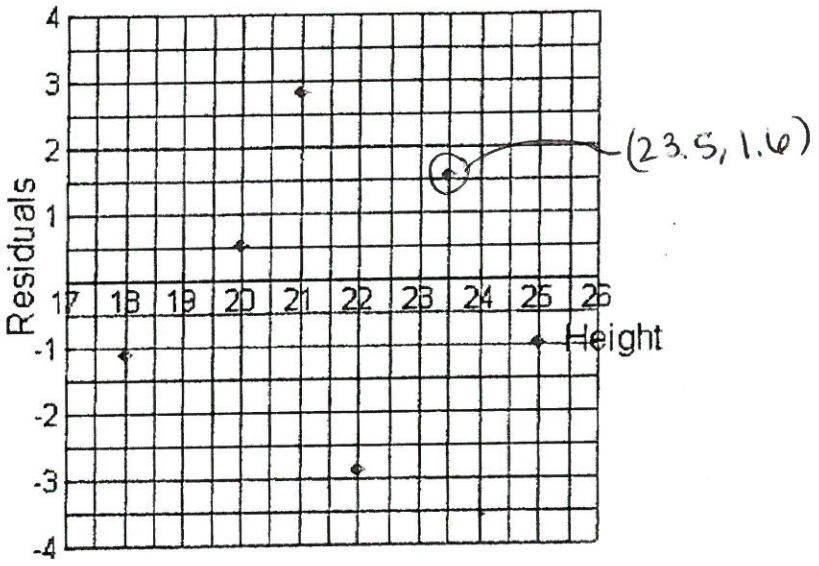
d. What is the value of the correlation coefficient for number of magazines published in the US and years since 1988? Interpret this correlation.

Since the slope is positive, the correlation coefficient is the positive square root of R^2.

$R^2 = 0.848$ $\quad r = +\sqrt{0.848} = 0.921$

Since the correlation coefficient $r = 0.921$, there is a strong, positive, linear relationship between the number of magazines published in the US and the year.

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?

Yes, a linear model is appropriate. The residual plot shows no pattern and a test for slope shows that there is a relationship ($H_0: \beta = 0$ $H_A: \beta > 0$ where β is the slope of the weight vs. height graph $df = 4$ $t = 8.987$ p-value = 0.00004 < 0.05)

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

$\widehat{\text{weight}} = -13.430 + 3.69516(\text{height})$

weight: in lbs

height: in inches

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.

Dakota's predicted weight is

$\hat{y} = (3.69516)(23.5) - 13.430 = 73.4$ lbs.

Dakota's residual (from the graph) is approximately 1.55 to 1.6

Dakota's actual weight is 75 lbs.

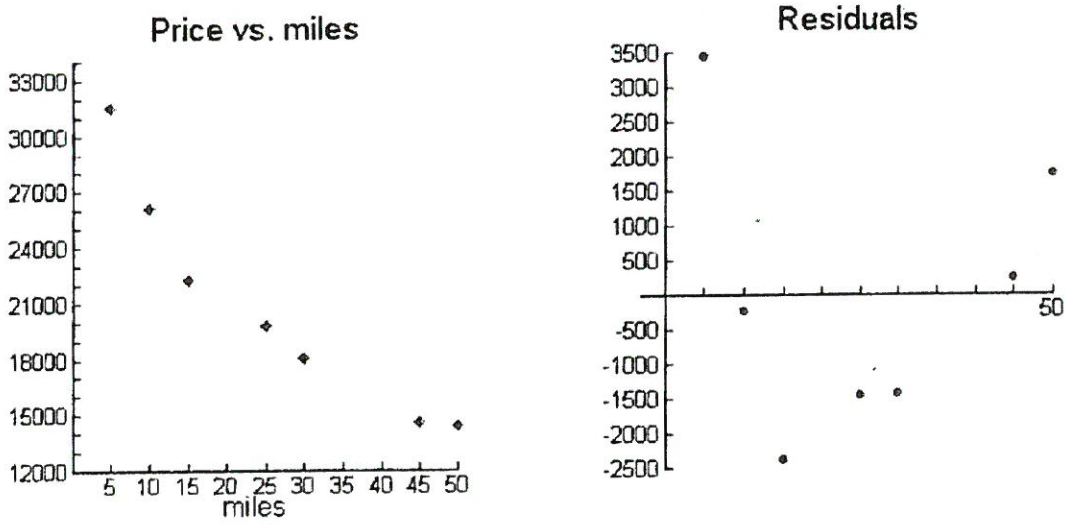
residual = observed − predicted

$1.6 = \text{observed} - 73.4$

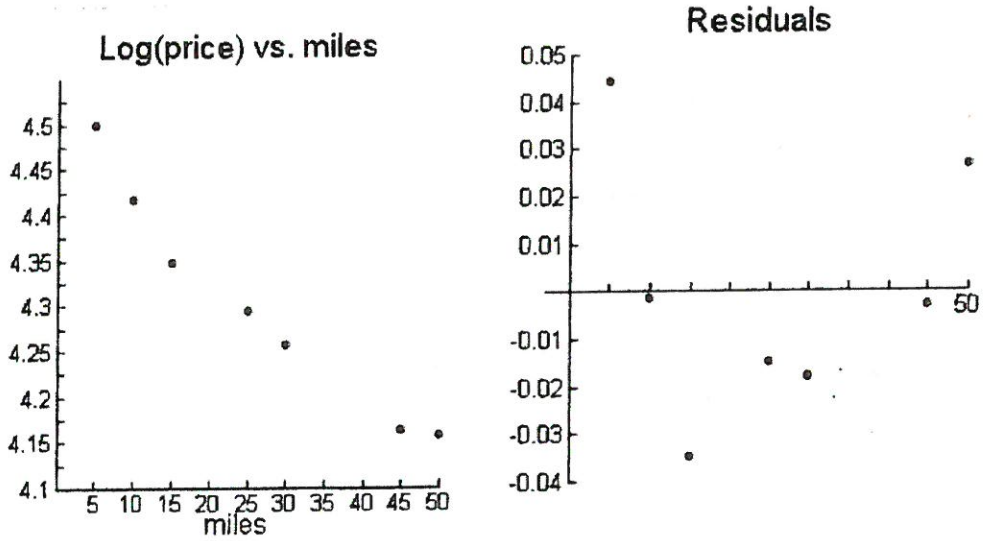
$\text{observed} = 1.6 + 73.4 = 75$ lbs

3. As more miles are driven in a car, the resale value of the car generally declines. This is called depreciation. For a certain make and model of car, information is gathered on the resale price in dollar and the number of miles driven (in thousands of miles). The scatterplot of price (y) versus miles (x), the residual plot, and the least squares regression line is shown for this data. In addition, the scatterplot, residual plot, and the accompanying best fit lines are shown for two other models using the common logarithm.

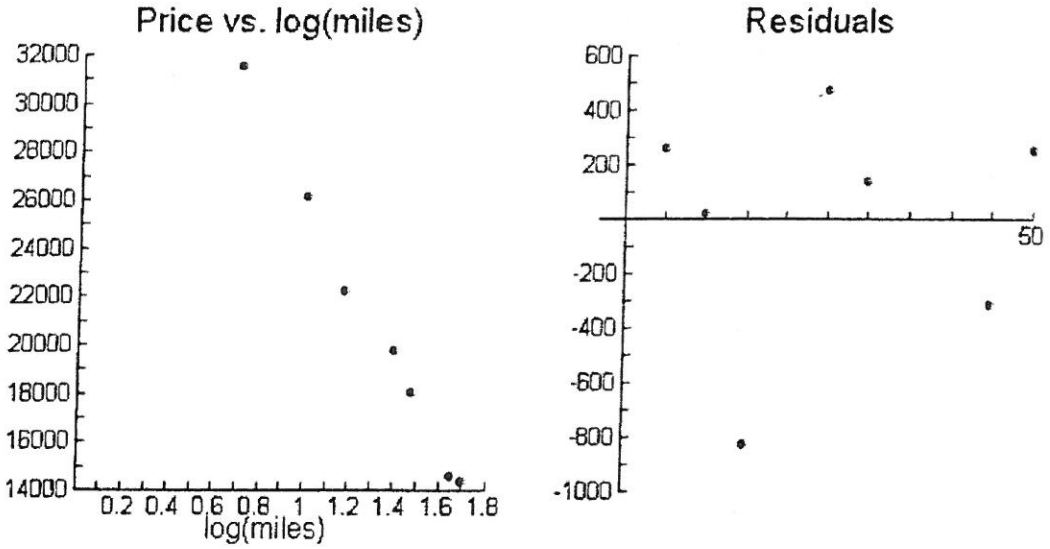
Model 1: $\hat{y} = 29784 - 343.58(x)$ r = -0.9452



Model 2: $\widehat{\log(y)} = 4.4901 - 0.0071910(x)$ r = -0.9765



Model 3: $\hat{y} = 43254 - 17153 \log(x)$ $r = -0.9975$



a. Using Model 1, estimate a resale price for a car of this make and model which has been driven 35,000 miles. $x = 35$

$\widehat{\text{resale price}} = 29784 - 343.58(35) = \$17758.70$

The estimated resale price for a car of this make and model which has been driven 35000 miles is $17758.70.

b. Model 1 is not the most appropriate to use to compute an estimated resale price. Explain why it is not appropriate, and determine whether Model 2 or Model 3 is better. The scatterplot slows a slight curve and the residual plot shows a pattern. Model 3 is the best because the scatterplot looks straightest and the residual plot has no pattern. Model 2 suffers from the same problems as model 1: a curved scatterplot and a residual plot with a pattern.

c. Use the model you chose in part (b) to estimate a resale price for a car of this make and model that has been driven 35,000 miles. $x = 35$

$\widehat{\text{resale price}} = 43254 - 17153 \cdot \log(35) = \$160768.60$

The estimated resale price for a car of this make and model which has been driven 35000 miles is $160768.60 based on model 3.