AP* Statistics Review

Linear Regression

Teacher Packet

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
  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)

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.

Transforming to get a linear model

- When a graph of y vs. x does not appear linear, 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, first substitute x into the equation, then perform the inverse operation to get the predicted y value.

Example: Bacteria are allowed to grow in a petri dish. The number of bacteria is recorded each hour. This data results in an exponential graph. However, if we take the logarithm of the y values, we can get the following linear model.

y = number of bacteria

x = time in hours

$$\log \hat{y} = 1.457 + 0.239x$$

Predict the number of bacteria after 6 hours.

Solution: $\log \hat{y} = 1.457 + (0.239)(6) = 2.891$

$$10^{\log \hat{y}} = 10^{2.891}$$

$$\hat{y} = 10^{2.891} = 778 \text{ bacteria}$$

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

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

8. Which is true?

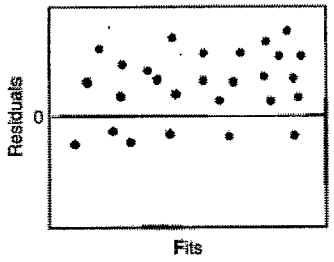
I. Random scatter in the residuals indicates a linear model.
II. If two variables are very strongly associated, then the correlation between them will be near +1.0 or −1.0.
III. Changing the units of measurement for x or y changes the correlation coefficient.

(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.

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%		

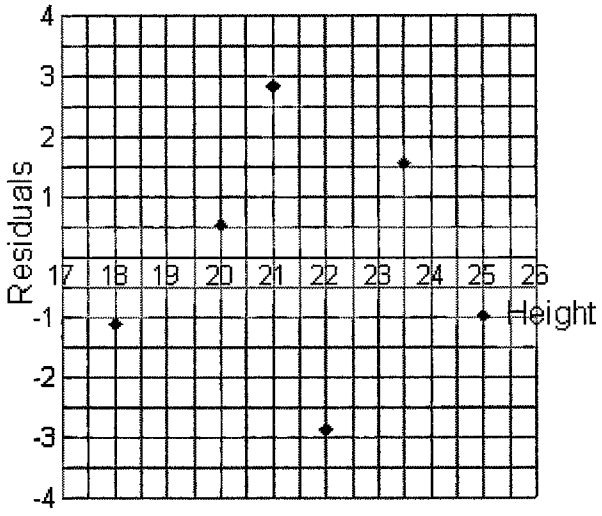
(a) What is the value of the slope of the least squares regression line? Interpret the slope in the context of this situation.

(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.

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%		

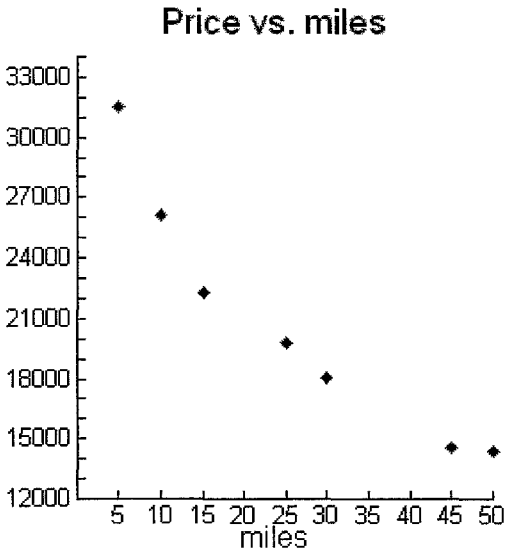
(a) Is a line an appropriate model to use for these data? What information tells you this?

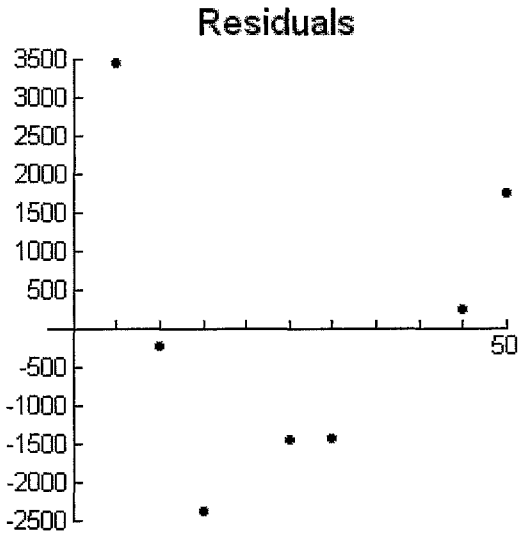
(b) Write the equation of the least squares regression line. Identify any variables used in this equation.

(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.

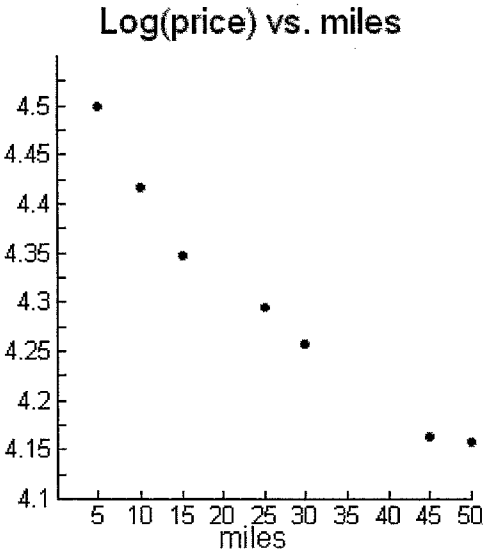
3. As more miles are driven in a car, the resale value of the car generally declines. This is called depreciation. For a certain make and model of car, information is gathered on the resale price in dollars and the number of miles driven (in thousands of miles). The scatterplot of price (y) versus miles (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.

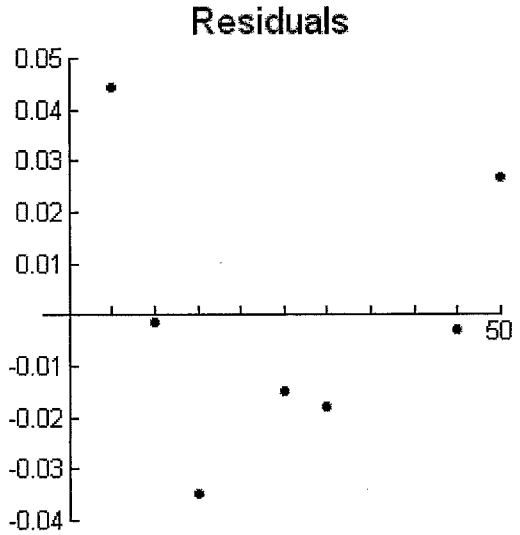
Model 1: $\hat{y} = 29784 - 343.58x$ $r = -0.9452$



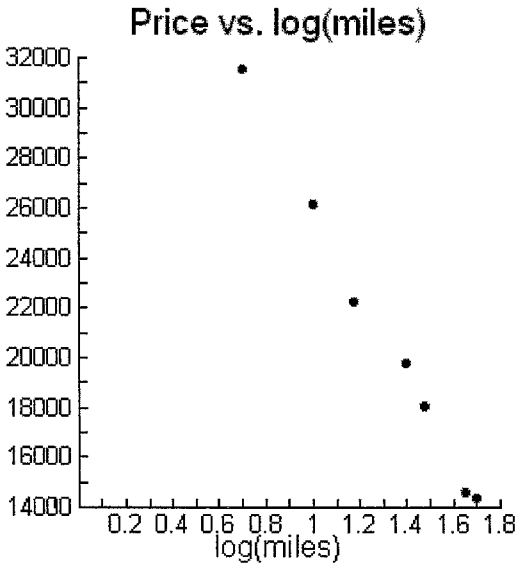


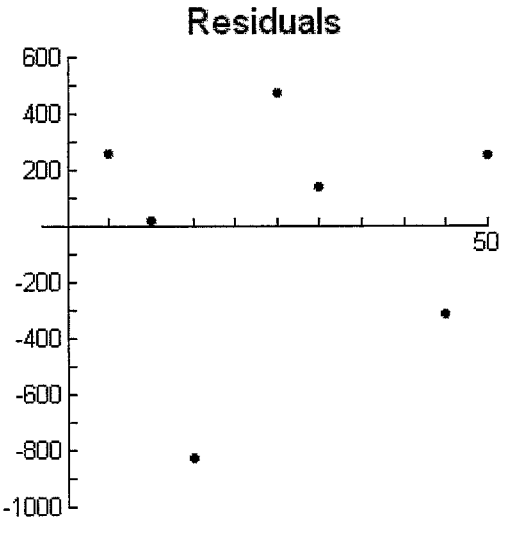
Model 2: $\log \hat{y} = 4.4901 - .0071910x$ $r = -0.9765$





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.

(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.

(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.

Key to Linear Regression Multiple Choice

1. C — Distractors: B is r^2; D is outlier; E is correlation coefficient r
2. B — $\hat{y} = -2.29 + (1.70)(5) = 6.21$; residual $= 6 - 6.21 = -0.21$
3. C — Tall fathers generally have tall sons, so r should be positive. But r cannot be over 1, and r will not be exactly 1 unless every data point falls on the best fit line, which should not be true for this data.
4. A — slope $= r\frac{s_y}{s_x} = (.63)\frac{1.9}{2.5} = .4788$
5. E — Distractors: A and C switch slope and intercept; B treats slope as a % instead of the ratio of \$1000/year
6. D — Distractors: A and E include categorical variables instead of numerical variables. In C, r cannot be over 1. In B, r should not have units.
7. E — $\ln \hat{h} = -1.2 + 1.4\ln(60) = 4.532$; $\hat{h} = e^{4.532} = 92.95$ ft
8. C
9. A — $r = \pm\sqrt{.64} = \pm 0.80$; Without the graph the direction of the association cannot be determined.
10. B — Too many of the residuals are positive; since residuals are $y - \hat{y}$, that means the actual values are larger than the predicted values.

Rubric for Linear Regression Free Response

1. **Solution**

Part (a):

The slope is 325.39 magazines per year. For each year since 1988, the number of magazines published in the US increases by about 325, on average.

Part (b):

The y-intercept is 13549.9 magazines. The predicted number of magazines published in the US in 1988 (year 0) is 13550 magazines.

Part (c):

1999 is year 11 (because 1999-1988 = 11).

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

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

Part (d):

Since the slope is positive, the correlation coefficient is the positive square root of 0.848: $r = +\sqrt{0.848} = 0.921$

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

Scoring

All parts can be essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is correct if 1) the numerical value is correct, 2) correct units are given for the slope, 3) the interpretation is correct and in context, and 4) the interpretation distinguishes between the model and the data by using words like about, approximately, or on average. The slope value may be rounded in the interpretation.

Part (a) is partially correct if the student correctly does 2 or 3 of the items listed above.

Part (a) is incorrect if the student correctly does 0 or 1 of the items listed above.

Part (b) is correct if 1) the numerical value is correct, 2) correct units are given for the y-intercept, 3) the interpretation is correct and in context, and 4) the interpretation distinguishes between the model and the data by using words like about, approximately, or predicted. The y-intercept value may be rounded in the interpretation.

Part (b) is partially correct if the student correctly does 2 or 3 of the items listed above.

Part (b) is incorrect if the student correctly does 0 or 1 of the items listed above.

Part (c) is essentially correct if the student identifies 1999 as year 11 and uses that value in a correct regression equation to find the number of magazines in 1999.

Part (c) is partially correct if the student correctly identifies 1999 as year 11 but doesn't use that value in a correct regression equation
OR
Has a correct regression equation but uses the wrong year

Part (d) is essentially correct if the value for r is correct and the interpretation is correct and in context.

Part (d) is partially correct if only one of the value for r or the interpretation in context is correct.

Each essentially correct response is worth 1 point; each partially correct response is worth half a point.

4 Complete Response

3 Substantial Response

2 Developing Response

1 Minimal Response

If a response is between two scores (for example, 2.5 points), use a holistic approach to determine whether to score up or down depending on the strength of the response and communication.

2. **Solution**

Part (a):

Yes, a linear model is appropriate. The residual plot shows no pattern and a test for slope shows that there is a relationship ($H_0 : \beta_1 = 0$; $H_a : \beta_1 > 0$; where β_1 is the slope of the weight vs. height graph, $df = 4$, $t = 8.987$; p-value $= 0.0004$).

Part (b):

$\hat{y} = 3.6956x - 13.430$ where x = height in inches and $\hat{y}$ = predicted height in pounds

Part (c):

Dakota's predicted weight is $\hat{y} = (3.6956)(23.5) - 13.430 = 73.4$ pounds.

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

$$\text{residual} = y - \hat{y}$$
$$1.6 = y - 73.4$$
$$y = 73.4 + 1.6 = 75 \text{ pounds}$$

Dakota's actual weight is 75 pounds.

Scoring

Part (a) can be essentially correct (E) or incorrect (I). Parts (b) and (c) can be essentially correct (E), partially correct (P), or incorrect (I).

Part (a) can be essentially correct even if it fails to mention the linear regression t test as long as the residual graph is discussed.

Part (a) is incorrect if the only evidence for linearity given is the value of the correlation coefficient, $r = 0.976$.

Part (b) is essentially correct if the correct numerical values of both the slope and y-intercept are present in the equation **and** both variables are identified. Note: variable names (height and weight) may be used in the equation in place of x and y for full credit.

Part (b) is partially correct if the correct numerical values of both the slope and y-intercept are present in the equation, but the variables are not identified

OR

both variables are identified, but the numerical values of the slope and y-intercept are incorrect

OR

Only one numerical value is correct and only one variable is identified.

Part (c) is essentially correct if 1) the predicted weight is correct, 2) an appropriate residual (between 1.5 and 1.7) is read from the graph, and 3) the actual weight is computed correctly with work shown. The weights must be distinguished by labels (actual, predicted) or symbols (y for actual weight, $\hat{y}$ for predicted weight).

Part (c) is partially correct if two of the three tasks are completed correctly. The computation of the actual weight can be considered to be correct if an incorrect residual is substituted correctly into the residual formula.

Pact (c) is incorrect if zero or one of the tasks is completed correctly.

4 Complete Response

All parts essentially correct.

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

OR

One part essentially correct and two parts partially correct

1 Minimal Response

One part essentially correct and either zero or one part partially correct

OR

No parts essentially correct and two parts partially correct

3. **Solution**

Part (a):

Using Model 1 gives an estimated resale price of $17,758.70:

$$\hat{y} = 29784 - (343.58)(35) = \$17758.70$$

Part (b):

Model 1 is not appropriate because the scatterplot shows a slight curve and the residual plot shows a pattern. Model 3 is 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.)

Part (c):

Using Model 3 gives an estimated resale price of $16,768.60:

$$\hat{y} = 43254 - (17153)\log 35 = \$16768.60$$

Scoring

Parts (a) and (c) can be essentially correct (E) or incorrect (I). Part (b) can be essentially correct (E), partially correct (P), or incorrect (I).

Part (a) is essentially correct if the correct answer is given and work is shown.
Part (a) is incorrect if the answer is wrong OR if the answer is correct but no work is shown.

Part (b) is essentially correct if the student discusses the pattern in the residual plot as a shortcoming AND chooses Model 3 because its residual plot is scattered.
Part (b) is partially correct if the student correctly does only one of the above.

Part (c) is essentially correct if the student makes a correct prediction based on the model chosen in part (b) and work is shown. Note: If Model 2 is chosen, the predicted resale price is $17,314.70:

$$\log \hat{y} = 4.4901 - (.0071910)(35) = 4.23841$$

$$\hat{y} = 10^{4.238415} = \$17314.70$$

4 Complete Response

All parts essentially correct.

3 Substantial Response

Two parts essentially correct and one part partially correct

2 Developing Response

Two parts essentially correct and no parts partially correct

1 Minimal Response

One part essentially correct and either zero or one part partially correct