NATIONAL MATH + SCIENCE INITIATIVE

AP Statistics

Linear Regression

Student Handout

2016-2017 EDITION

Linear Regression

Linear Regression t-test Test Notes

2002B FR Q1 (extended)

1. Animal-waste lagoons and spray fields near aquatic environments may significantly degrade water quality and endanger health. The National Atmospheric Deposition Program has monitored the atmospheric ammonia at swine farms since 1978. The data on the swine population size (in thousands) and atmospheric ammonia (in parts per million) for one decade are given below.

Year	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Swine Population	0.38	0.50	0.60	0.75	0.95	1.20	1.40	1.65	1.80	1.85
Atmospheric Ammonia	0.13	0.21	0.29	0.22	0.19	0.26	0.36	0.37	0.33	0.38

(a) Construct a scatterplot for these data.

(b) The value for the correlation coefficient for these data is 0.85. Interpret this value.

(c) Based on the scatterplot in part (a) and the value of the correlation coefficient in part (b), does it appear that the amount of atmospheric ammonia is linearly related to the swine population size? Explain.

(d) What percent of the variability in atmospheric ammonia can be explained by swine population size?

e) What is the equation of the least squares regression line? Interpret the slope and y-intercept.

f) How would we check to make sure a linear model is really the best model?

f) What does “least squares regression equation” mean?

g) A population of 9,500 pigs gives a residual of −.063. What is the actual ammonia level?

h) What is the residual for the point (1.5 , .30)?

2. A study concerning the relationship between the average amount of time spent outside per month y (in hours) and the population of a city x (in millions) resulted in the following MINITAB output.

Predictor	Coef	Stdev	t-ratio	P
Constant	17.15	2.395	3.71	0.002
popn	–6.650	1.910	8.72	0.016

s = 5.454 R-sq = 84.4% R-sq(adj) = 83.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	1	2260.5	2260.5	76.00	0.000
Error	14	416.4	29.7		
Total	15	2676.9			

a) What is the equation of the least squares regression line?

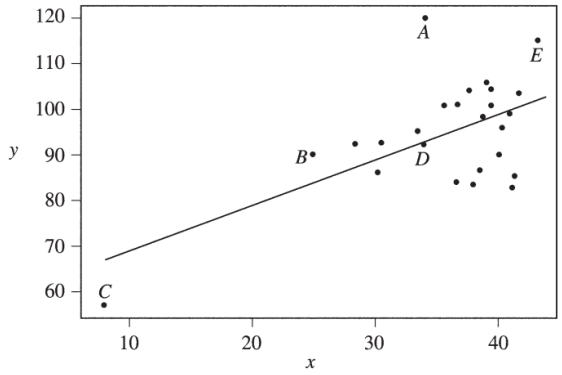
b) Estimate the mean amount of time spent outside for a city having a population of 1 million people.

c) Determine the proportion of the observed variation in time spent outside that can be attributed to the population of a city.

d) What is the correlation coefficient?

Multiple Choice

Questions 1 and 2 refer to the following scatterplot.



1. In the scatterplot of y versus x shown above, the least squares regression line is superimposed on the plot. Which of the following points has the largest residual?

   A) A
   B) B
   C) C
   D) D
   E) E

2. Which of the following points has the greatest influence on the strength of the correlation coefficient?

   A) A
   B) B
   C) C
   D) D
   E) E

3. There is a linear relationship between the number of chirps made by the striped ground cricket and the air temperature. A least squares fit of some data collected by a biologist gives the model

$$\hat{y} = 25.2 + 3.3x \qquad 9 < x < 25$$

   where x is the number of chirps per minute and $\hat{y}$ is the predicted temperature in degrees Fahrenheit. What is the estimated increase in temperature that corresponds to an increase of 5 chirps per minute?

   A) 3.3° F
   B) 16.5° F
   C) 25.2° F
   D) 28.5° F
   E) 41.7° F

4. The equation of the least squares regression line for the points on a scatterplot (not pictured) is $\hat{y} = 2.3 + 0.37x$. What is the residual for the point (4, 7)?

   A) 3.22
   B) 3.78
   C) 4.00
   D) 5.52
   E) 7.00

5. The correlation between two scores *X* and *Y* equals 0.75. If both the *X* scores and the *Y* scores are converted to *z*-scores, then the correlation between the *z*-scores for *X* and the *z*-scores for *Y* would be

   A) –0.75
   B) –0.25
   C) 0.0
   D) 0.25
   E) 0.75

6. A least squares regression line was fitted to the weights (in pounds) versus age (in months) of a group of many young children. The equation of the line is

   $$\hat{y} = 16.6 + 0.65x$$

   where $\hat{y}$ is the predicted weight and *x* is the age of the child. The residual for the prediction of the weight of a 20-month-old child in this group is -4.60. Which of the following is the actual weight, in pounds, for this child?

   A) 13.61
   B) 20.40
   C) 25.00
   D) 29.60
   E) 34.20

7. A sports medicine surgeon is interested in the relationship between the range of motion of baseball pitchers and the number of years playing the sport. Based on collected data, the least squares regression line is $\hat{y} = 250.35 - 1.71x$, where x is the number of years the player has played professional baseball and y is the number of degrees of motion in the players pitching arm. Which of the following best describes the meaning of the slope of the least squares regression line?

   A) For each increase of one degree of motion, the estimated number of years played decreases by 1.71.
   B) For each increase of one degree of motion, the estimated number of years played increases by 1.71.
   C) For each increase of one year played, there is an estimated increase in degrees of motion of 1.71.
   D) For each increase of one year played, the number of degrees of motion decreases by 1.71.
   E) For each increase of one year played, there is an estimated decrease in degrees of motion of 1.71.

8. A real estate company is interested in developing a model to estimate the prices of homes in a particular area of a large metropolitan area. A random sample of 30 recent home sales in the area is taken, and for each sale, the size of the house (in square feet), and the sale price of the house (in thousands of dollars) is recorded. The regression output for a linear model is shown below.

Variable	**Coef**	**S.E. Coeff**	**t**	**p**
Constant	13.465	16.7278	0.805	0.4276
Size	0.123	0.00744	16.662	0.0000
$S = 16.3105$		R-sq = 0.908	R-sq(adj) = 0.905	

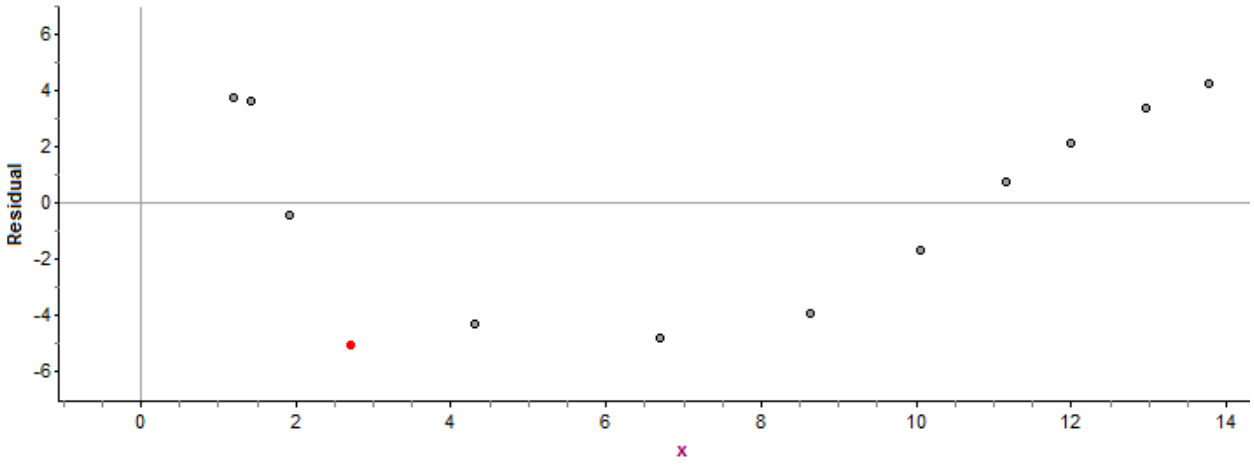
   What percent of the selling price of the home is explained by the linear relationship with size of the home?

   A) 82.4%
   B) 90.5%
   C) 90.8%
   D) 95.1%
   E) 95.3%

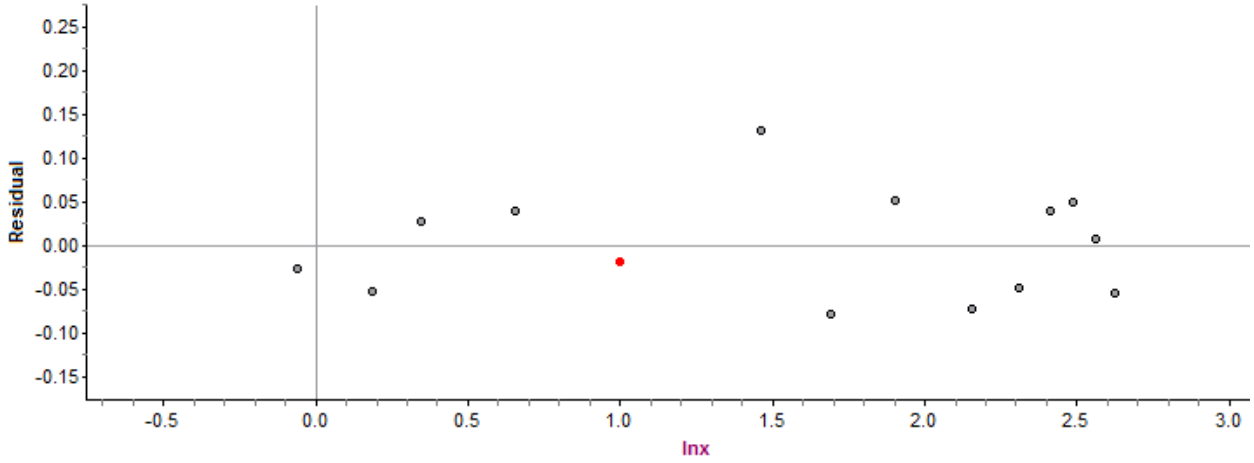
9. Using the information in Question 8 above, what is the equation of the least squares regression line?

   A) $\hat{y}$ = 13.465x + 16.7278
   B) $\hat{y}$ = 13.465x + 0.123
   C) $\hat{y}$ = 0.123x + 0.00744
   D) $\hat{y}$ = 0.123x + 13.465
   E) $\hat{y}$ = 16.3105x + 0.908

10. Two measures x and y were taken on 15 subjects. The first of two regressions, Regression I, yielded $\hat{y} = 30.72 - 2.01x$ and had the following residual plot.



The second regression, Regression II, yielded $\ln \hat{y} = 3.63 - 0.61 \ln x$ and had the following residual plot



Which of the following conclusions is best supported by the evidence above?

A) There is a linear relationship between x and y, and Regression I yields a better fit.
B) There is a linear relationship between x and y, and Regression II yields a better fit.
C) There is a positive correlation between x and y.
D) There is a nonlinear relationship between x and y, and Regression I yields a better fit.
E) There is a nonlinear relationship between x and y, and Regression II yields a better fit.

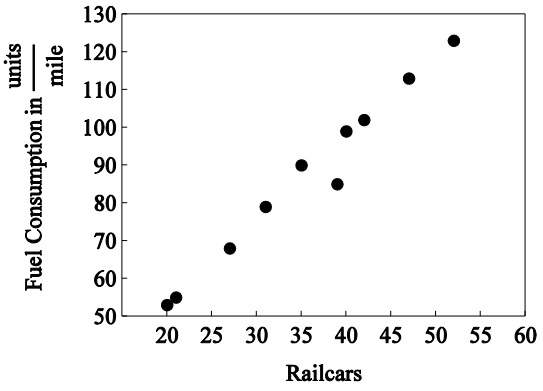
Additional Free Response Questions

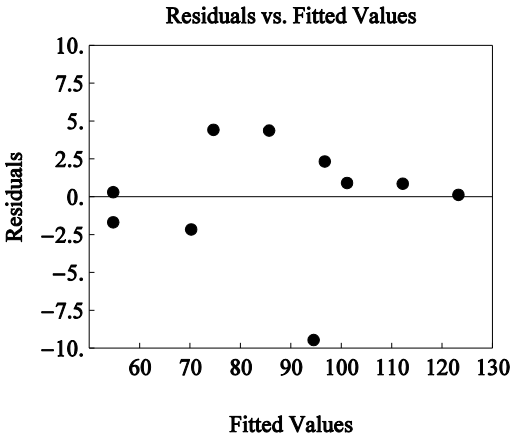
2005 Q3 revised

The Western Canadian Railroad is interested in studying how fuel consumption is related to the number of railcars for its trains on a certain route between Edmonton and Victoria Canada. A random sample of 10 trains on this route has yielded the data in the table below.

Number of Railcars	Fuel Consumption $\left(\frac{\text{units}}{\text{mile}}\right)$
21	55
20	53
35	90
31	79
47	113
42	102
39	85
52	123
40	99
27	68

A scatterplot, a residual plot, and the output from the regression analysis for these data are shown below.





Variable	Coef	S.E. Coeff	t	p
Constant	8.2689	4.9599	1.667	0.1340
Railcars	2.2093	0.1345	16.420	0.0000
$S = 4.226$		R-sq = 0.971		R-sq(adj) = 0.968

(a) Is a linear model appropriate for modeling these data? Clearly explain your reasoning.

(b) Suppose the fuel consumption cost is $42 per unit. Give a point estimate (single value) for the change in the average cost of fuel per mile for each additional railcar attached to a train. Show your work.

c) Interpret the value of r^2 in the context of this problem.

(d) Would it be reasonable to use the fitted regression equation to predict the fuel consumption for a train on this route if the train had 5 railcars? Explain.

(e) What is the value of the correlation coefficient? Interpret this value in context.

(f) What is the residual for the train with 40 cars? Interpret this value in context.

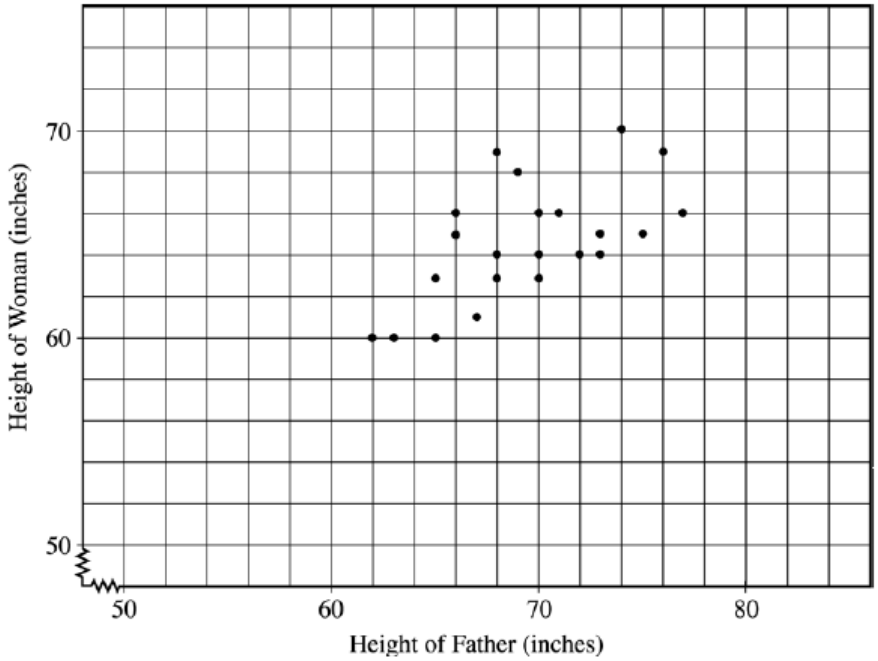
(g) Suppose the fuel consumption cost is $42 per unit. If the trip from Victoria to Edmonton is 775 miles, estimate the operating cost for a train with 33 cars to make the trip.

(h) Describe the effect of adding a train with 34 rail cars and a fuel consumption of 130 units/mile on the correlation coefficient. (no calculations are necessary)

(i) Describe the effect of adding a train with 34 rail cars and a fuel consumption 130 units/mile on the slope of the LSRL. (no calculations are necessary)

2007B Q4

Each of 25 adult women was asked to provide her own height (y), in inches, and the height (x), in inches, of her father. The scatterplot below displays the results. Only 22 of the 25 pairs are distinguishable because some of the (x, y) pairs were the same. The equation of the least squares regression line is $\hat{y} = 35.1 + 0.427x$.



(a) Draw the least squares regression line on the scatterplot above.

(b) One father's height was $x = 67$ inches and his daughter's height was $y = 61$ inches. Circle the point on the scatterplot above that represents this pair and draw the segment on the scatterplot that corresponds to the residual for it. Give a numerical value for the residual.

(c) Suppose that the point $x = 84$, $y = 71$ is added to the data set. Would the slope of the least squares regression line increase, decrease, or remain about the same? Explain.
(Note: No calculations are necessary to answer this question.)

Would the correlation increase, decrease, or remain about the same? Explain.
(Note: No calculations are necessary to answer this question.)

Exploring Bivariate Data Notes

Communication, skills, and understanding…

- Title, scale and label the horizontal and vertical axes
- Comment on the direction, shape (form), and strength of the relationship and unusual features (possible outliers) in context
- Include the "hat" on the *y*-variable and identify both variables in your least squares regression equation
- Interpret the *y*-intercept or slope in the context of the problem
    - The intercept provides an estimate for the value of *y* when *x* is zero.
    - The slope provides an estimated amount that the *y*-variable changes (or the amount that the *y*-variable changes on average) for each unit change in the *x*-variable.
- Residual = observed y – predicted y; Resid = $(y - \hat{y})$
- Examine a residual plot and make sure that the residuals are randomly scattered about the horizontal axis to determine whether or not the model is a good fit.
- Avoid using the least squares line to predict outside the domain of the observed values of the explanatory variable. Extrapolation is risky!
- An influential point is a point that noticeably affects the slope of the regression line when removed from (or added to) the data set. An outlier is a point that noticeably stands apart from the other points and has a large residual.
- The magnitude of the correlation coefficient provides information about the strength of the linear relationship between two quantitative variables over the observed domain.
- Interpretation of the correlation coefficient:
    - Comment on the strength using the magnitude of the correlation coefficient
        - If the value of *r* is close to 1 or -1; there is a strong linear association
        - If the value of *r* is close to 0, there is a very weak linear association and could suggest a strong non-linear relationship.
        - The magnitude does <u>not</u> provide information about whether a linear model is appropriate. You must also consider the residual plot.
    - Comment on the direction of the linear relationship in context.
    - Correlation does not imply causation.
- The percent of variation in the observed *y*-values that can be attributed to the linear relationship with the *x*-variable is r^2, coefficient of determination. Interpret r^2 in context.

Calculator Use

You may need to use your calculator to create a scatter plot, compute the equation of a least-squares regression line (and the values of *r* and r^2), graph the regression line with the data, and create a residual plot. Generally, computer output and graphs are provided with bivariate data analysis questions, but you cannot be sure that these will be provided.