AP Statistics: Linear Regression Review – For each of the data sets, perform a complete bivariate analysis by performing all of the following steps:

a) Make a scatterplot on your calculator and sketch it on your paper.
b) Determine the association (pos or neg) and describe it in a sentence.
c) Find the equation of the least squares linear regression line.
d) Describe the meaning of the slope of the line in a sentence.
e) Find the correlation coefficient. What does it tell you about the data?
f) What is r^2 and what does it tell you about the data?
g) Make a residual plot on your calculator and sketch it on paper. Interpret the plot.
h) Are there any outliers/influential points? If so, remove them and recalculate the regression equation. What effect was there on the correlation and regression equation?

---

1. The following are median heights, in inches, of boys age 2 to 14.

Age (yr)	Median Height (in)
2	35.1
3	38.7
4	41.3
5	44.1
6	46.5
7	48.6
8	51.7
9	53.7
10	56.1
11	59.5
12	61.2
13	62.9
14	63.6

2. The following are population (in millions) and number of police officers (in thousands) for various states.

State	Pop	Police
CA	30.4	86.2
CO	3.4	9.2
FL	13.2	45.0
IL	11.5	39.9
IA	2.8	6.0
LA	4.2	11.8
ME	1.2	2.9
MS	5.2	14.6
NJ	7.7	30.5
TN	5.0	12.3
TX	17.3	46.2
VA	6.3	15.2
WA	5.0	10.9

3. The following are data on the number of online educational journals from 91-97.

Year	Journals
91	27
92	36
93	45
94	181
95	306
96	1093
97	2459

4. The following are data gathered by anthropologists digging in Mayan ruins. X represents the depth of the soil in centimeters and Y is the percent of the mineral *montmorillonite* in the soil.

Depth	% montmorillonite
40	58
50	34
60	32
70	30
80	28
90	27
100	22

5. The following are data on various fast food hamburgers for various states. X represents the grams of fat in the burger and Y is the number of calories.

Fat	Calories
19	410
31	580
34	590
35	570
39	640
39	680
43	660

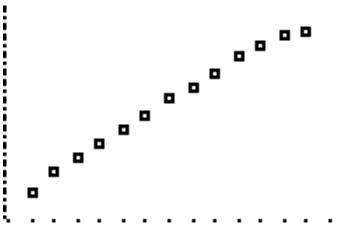
6. The following are data on the number of dollars (in millions) that the U.S. federal government spent on mathematics research from 1980 to 1990.

Year	Research $ (in millions)
80	91
81	118
82	128
83	134
84	151
85	184
86	185
87	205
88	212
89	230
90	245

AP Statistics: Linear Regression Review – Answer Key

1. x = age in years, y = median height for boys age 2-14, in inches

   a. Scatter Plot



   b. Positive Association – As age increases, height increases

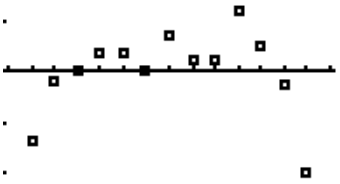
   c. y = 31.567 + 2.429x OR height = 31.567 + 2.429age

   d. On average, boys age 2-14 grow 2.429 inches per year

   e. r = .996. which tells us that there is a strong relationship between age and height

   f. r^2 = 99%, so 99% of the change in height can be attributed to age

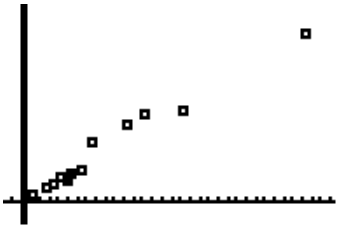
   g. Residual Plot is random, so a line is a good model for data



   h. There are no outliers

2. x = state population in millions, y = number of police officers in 1000's

   a. Scatter Plot



   b. Positive Association – As pop. increases, number of police officers increases

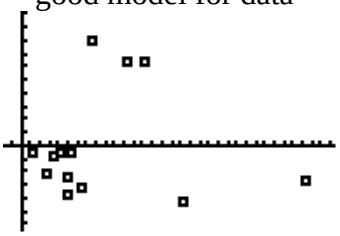
   c. y = -.0008 + 2.921x OR cops = -.0008 + 2.921pop

   d. On average, there are 2,921 police officers per 1 million people

   e. r = .984. which tells us that there is a strong relationship between pop and police

   f. r^2 = 97%, so 97% of the change in police can be attributed to pop

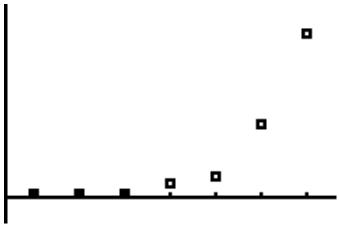
   g. Residual Plot is random, so a line is a good model for data



   h. CA is an outlier. Removal causes r to get lower, equation also changes

3. x = year, y = # of online education journals

   a. Scatter Plot



   b. Positive Association – As time increases, number of journals increases

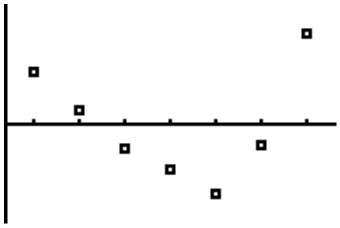
   c. y = -31874.5 + 345.4x OR journals= -31874.5 + 345.4year

   d. On average, the number of journals increases 345.4 per year

   e. r = .824. which tells us that there is a moderate relationship between time and # of journals

   f. r^2 = 68%, so 68% of the change in # of journals can be attributed to time

   g. Residual Plot shows a pattern, so a line is not the best model for data

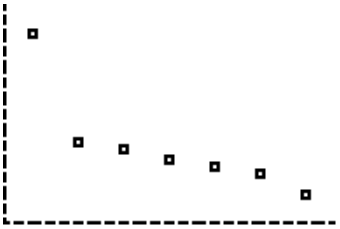


   h. There are no outliers

AP Statistics: Linear Regression Review – Answer Key

4. x = soil depth in cm, y = % of *montmorillonite* in the soil

   a. Scatter Plot

   b. Negative Association – As soil depth increases, % *mont*… decreases

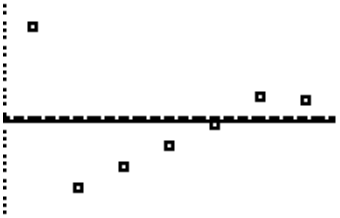
   c. y = 64.5 -.45x OR percent = 64.5 - .45depth

   d. On average, the percent decreases by .45% per cm of soil depth

   e. r = -.833. which tells us that there is a moderate relationship between depth and percent

   f. r^2 = 69%, so 69% of the change in percent can be attributed to soil depth

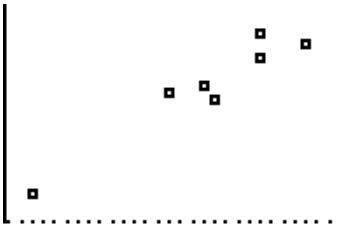
   g. Residual Plot shows a pattern, so a line may not be the best model

   h. (40,58) is an outlier. Removal causes r to get much better (-.98), equation also changes

5. x = fat content in grams, y = calories in burger

   a. Scatter Plot

   b. Positive Association – As fat increases, calories increase

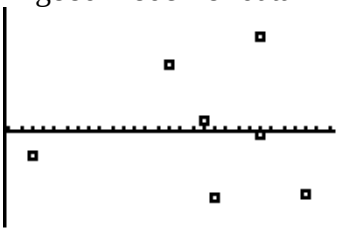
   c. y = 210.95 + 11.056x OR calories = 210.95 + 11.056fat

   d. On average, calories increase 11.056 per gram of fat in burgers

   e. r = .961. which tells us that there is a strong relationship between fat and calories

   f. r^2 = 92%, so 92% of the change in calories can be attributed to fat

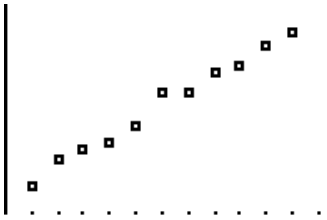
   g. Residual Plot is random, so a line is a good model for data

   h. (19,410) is an outlier. Removal causes r to get lower (.836), equation also changes slightly (the slope)

6. x = year, y = research funding (in millions)

   a. Scatter Plot

   b. Positive Association – As time increases, amount of research funding increases

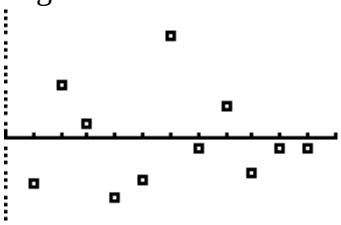
   c. y = -1100.73 + 14.96x OR funding= -1100.73 + 14.96year

   d. On average, the amount of funding increases $14.96 million per year

   e. r = .993. which tells us that there is a very strong relationship between time and research funding

   f. r^2 = 98.6%, so 98.6% of the change in research funding can be attributed to time

   g. Residual Plot is random, so a line is a good model for data

   h. There are no outliers