Mr Murphy **HW Pg 236 #5.36, 5.37, 5.39, 5.47**
AP Statistics
4.3 Assessing Fit of a Line

#Goals:

1. Calculate residuals and make a residual plot.
2. Determine if regression model is a good fit for data.
3. Interpret, in context, the correlation coefficient and the coefficient of determination.
4. Interpret s_e.
5. Distinguish between an outlier and influential point in a regression setting.

- Each residual is the difference between an observed *y* value and the corresponding predicted *y* value.

Residual = Actual *y* value - Predicted *y* value

- A **residual plot** is a scatterplot of the (*x*, residual pairs)
- A residual plot that has a *scattered pattern* indicates a good fitting model.
- A residual plot with a pattern suggests there is a better regression model.

Ex1 Consider the data on x = height (in inches) and y = weight (in pounds) for American females, age 30 - 39.

x	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
y	113	115	118	121	124	128	131	134	137	141	145	150	153	159	164

Find the LSRL equation and the correlation coefficient.
Calculate the residual for a height of 63 inches.

Calculate the residual for a height of 70 inches.

Use your calculator to find residuals - there are two ways.
$L_3 = L_2 - Y_1(L_1)$ -> StatPlot -> Scatterplot -> L_1,L_3
StatPlot -> Scatterplot -> L_1,Resid
Note: You find Resid with 2nd -> Stat -> Resid
Construct a residuals plot.

Determining if Model is a Good Fit for Data

Four factors go into your decision:

1. Does the scatterplot look linear?
2. r - should be close to +1 or -1 for a good fit
3. Residual plot - should be scattered for a good fit
4. s_e (the standard deviation of the residuals i.e. the average residual size) - should be small for a good fit

Is this regression equation found in Ex1 a good fit for the data?

Ex2 One measure of success of knee surgery is post-surgical range of motion for the knee joint. Post-surgical range of motion was recorded for 12 patients who had surgery following a knee dislocation. The age of each patient was also recorded.

The data are given in the chart.
Here is the partial MINITAB output:

Regression Analysis

The regression equation is

Range of motion = 108 + 0.871(Age)

Predictor	Coef	StDev	T	P
Constant	107.58	11.12	9.67	0.000
Age	0.8710	0.4146	2.10	0.062

s = 10.42 R-Sq = 30.6% R-Sq(adj) = 23.7%

Find the equation for the LSRL.
Calculate the residual for Patient 9.
Is the LSRL a reasonable fit for the data?

Patient	Age (x)	Range of Motion (y)
1	35	154
2	24	142
3	40	137
4	31	133
5	28	122
6	25	126
7	26	135
8	16	135
9	14	108
10	20	120
11	21	127
12	30	122

Ex3 If an observed y-value is below a line of best fit, then the residual is:

(a) positive
(b) negative
(c) equal to the squared residual
(d) greater than 1
(e) None of the above

- The **coefficient of determination**, r^2, is the fraction of the variation in the values of *y* that is explained by least-squares regression of *y* on *x*.

- Interpretation for r^2:
  _____% of the variation in the *y* variables can be explained by the *x* variables
  OR
  _____% of the variation in the *y* variables can be explained by the LSRL of *y* on *x*.

Ex4 Consider a data set A: (2, 8), (3, 6), (4, 9) and (5, 9). What is the value of r^2?

(a) -0.50 (b) 0.25 (c) 0.30 (d) 0.55 (e) 0.60

Interpret this value.

Ex5 The heart disease death rates per 100,000 people in the United States for certain years, as reported by the National Center for Health Statistics, were

Year:	1950	1960	1970	1975	1980
Death rate:	307.6	286.2	253.6	217.8	202.0

Which of the following is the correct interpretation of the coefficient of determination?

(a) The heart disease rate per 100,000 people has been dropping on average of 3.627 per year.
(b) The baseline heart disease rate is 7386.87
(c) The regression line explains 96.28% of the variation in heart disease rates over the years.
(d) The regression explains 98.12% of the variation in heart disease rates over the years.
(e) Heart disease will be cured in the year 2036.

- The **standard deviation about the LSRL** is denoted by s_e. You can think of this number as the standard deviation for the residuals. You want s_e to be small. Even if r^2 is high, you want s_e to be small. You can find s_e in a MINITAB output.

Regression Analysis

The regression equation is

Range of motion = 108 + 0.871(Age)

Predictor	Coef	StDev	T	P
Constant	107.58	11.12	9.67	0.000
Age	0.8710	0.4146	2.10	0.062

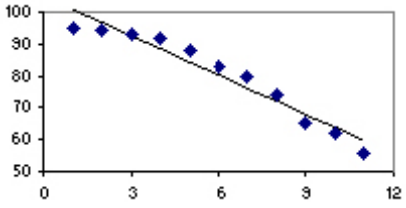
s = 10.42 R-Sq = 30.6% R-Sq(adj) = 23.7%

- An **outlier** is a data point separated from the rest of a set, i.e. an observation with a large residual.
- An observation that causes the value of the slope or the *y* intercept in the LSRL to be considerably different from what it would be had the observation been removed is **influential**.

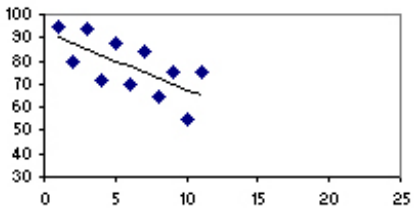
Outliers vs. Influential Points

Outliers greatly effect r^2, where influential points have more effect on the slope.

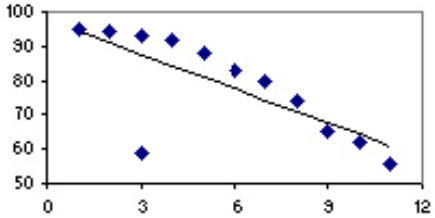
Regression equation: $\hat{y}$ = 104.78 - 4.10*x*
Coefficient of determination: r^2 = 0.94



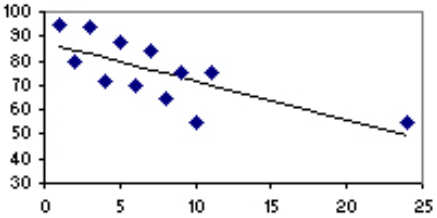
Regression equation: $\hat{y}$ = 92.54 - 2.5*x*
Slope: b_0 = -2.5
Coefficient of determination: r^2 = 0.46



Regression equation: $\hat{y}$ = 97.51 - 3.32*x*
Coefficient of determination: r^2 = 0.55

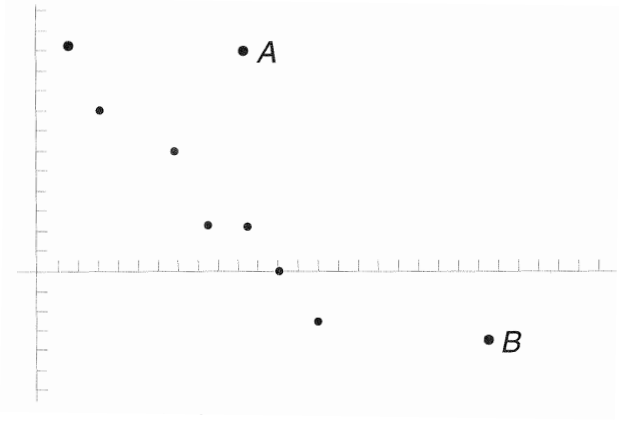


Regression equation: $\hat{y}$ = 87.59 - 1.6*x*
Slope: b_0 = -1.6
Coefficient of determination: r^2 = 0.52



Checkpoint:
Multiple Choice

1. In the graph below how would the regression line computed with point B (not including point A), differ from the regression line using the original data points (excluding points A and B)?



(a) The y intercept of the line with point B would be greater.
(b) The r^2 value of the line with point B would be larger.
(c) The slope of the line with point B would be less steep than the slope of the line without point B.
(d) The y intercept of the line with point B would be zero.
(e) None of the above.

2. A study found correlation $r = 0.61$ between the sex of a worker and his or her income. You conclude that

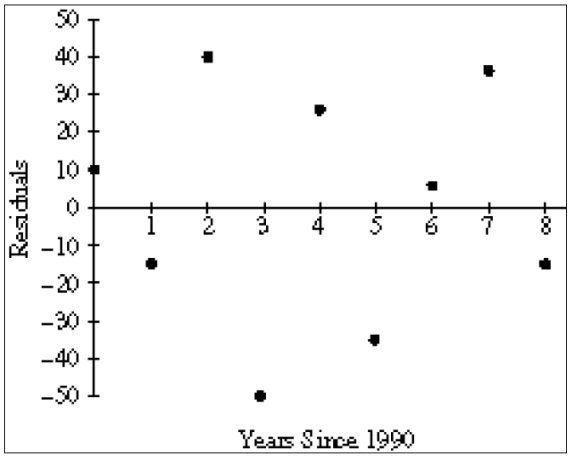
(a) women earn more than men on the average.
(b) women earn less than men on the average.
(c) an arithmetic mistake was made; this is not a possible value of r.
(d) this is nonsense because r makes no sense here.
(e) the correlation should have been $r = -0.61$.

3. Which of the following statements is/are true?
I. Correlation and regression require explanatory and response variables.
II. Scatterplots require that both variables be quantitative.
III.Every LRSL passes through $(\overline{x}, \overline{y})$.

(a) I and II only
(b) I and III only
(c) II and III only
(d) I, II, and III
(e) None of the above

Free Response

1. (1999 Q1) Lydia and Bob were searching the internet to find information on air travel in the United States. They found data on the number of commercial aircraft in the United States during the years 1990 - 1998. The dates were recorded as years since 1990. Thus, the year 1990 was recorded as year 0. They fit a least squares regression line to the data. The graph of the residuals and part of the computer output for their regression are given below.



Predictor	Coef	Stdev	t-ratio	p
Constant	2939.93	20.55	143.09	0.000
Years	233.517	4.316	54.11	0.000
s = 33.43				

(a) Is a line an appropriate model to use for these data? What information tells you this?
(b) What is the value of the slope of the LSRL? Interpret the slope in the context of this situation.
(c) What is the value of the intercept of the LSRL? Interpret the intercept in the context of this situation.
(d) What is the predicted number of commercial aircraft flying in 1992?
(e) What was the actual number of commercial aircraft flying in 1992?