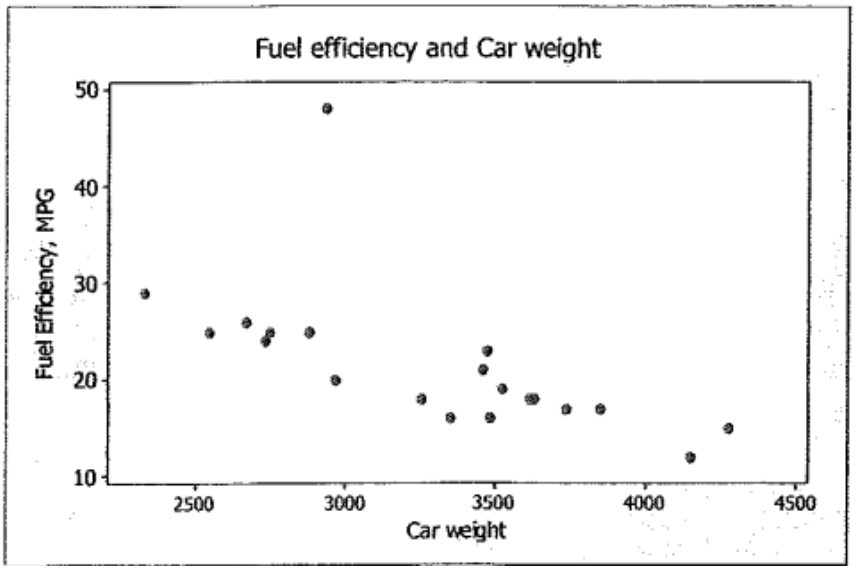
AP Statistics Ch 7,8,9 - Required Practice Name: _____Per: ____

The scatterplot below shows the fuel efficiency (in miles per gallon) and weight (in pounds) of twenty 2009 subcompact cars.



#1. Is there a clear explanatory variable and response variable in this setting? If so, tell which is which. If not, explain why not.

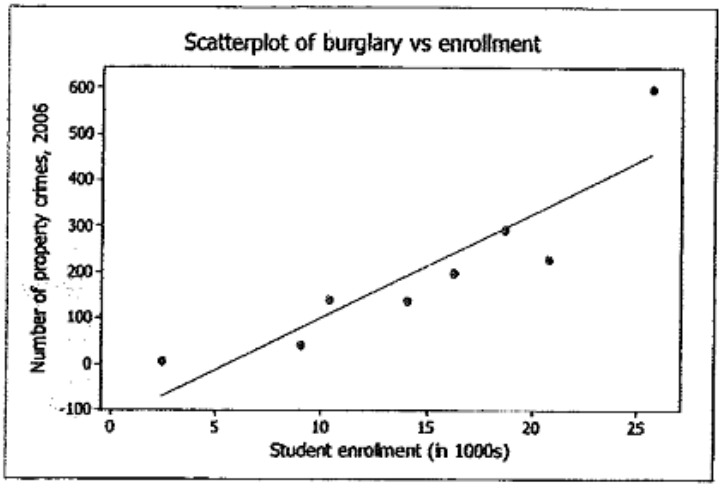
#2. Describe the association between fuel efficiency and car weight.

#3. Which of the following is closest to the correlation between car weight and fuel efficiency for these 20 vehicles? Explain.

r = -0.9 r = -0.6 r = 0 r = 0.4

The table and scatterplot below show the relationship between student enrollment (in thousands) and total number of property crimes (burglary and theft) in 2006 for eight colleges and universities in a certain U.S. state.

Enrollment (in 1000s) (x)	No. of Property Crimes (y)
16	201
2	6
9	42
10	141
14	138
26	601
21	230
19	294



#4. Use a calculator to find the equation of the least-squares regression line (LSRL).

#5. Interpret the slope of the LSRL in the context of the problem.

#6. How many crimes would you predict on a campus with enrollment of 14 thousand students? (Show your work)

#7. Find the residual for the campus with 14 thousand students and 138 property crimes. (Show your work). Interpret the value of the residual in the context of the problem.

#8. The value of r^2 for these data is 0.801. Interpret this value in the context of the problem.

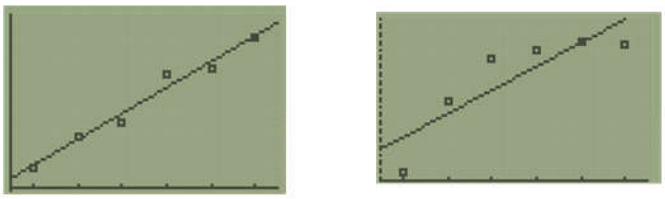
Residual Plots

A **residual plot** for a given linear regression shows the residual (r) vs. x.

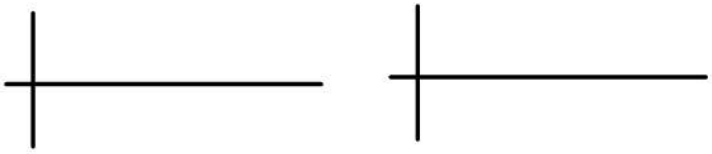
Examples: Sketch residual plots by hand for each data set

x	y
1	8
2	22
3	28
4	48
5	51
6	64

x	y
1	2
2	18
3	28
4	30
5	32
6	31



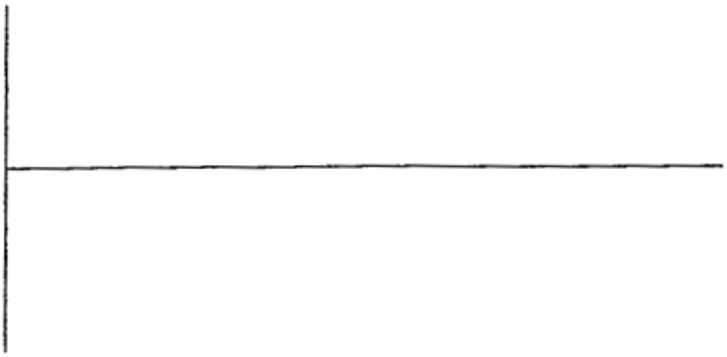
Scatter plots

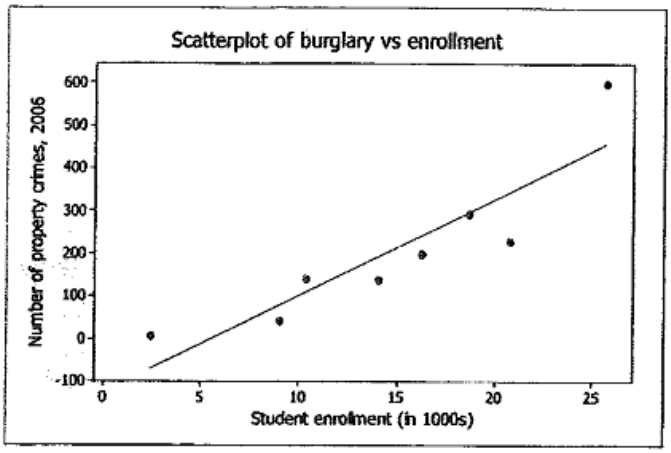


Residual plots

If the residuals are randomly scattered around '0' then you know that a linear model is appropriate. (Residuals make it easier to see non-linearity compared to scatterplots.)

#9. Use the scatterplot to make a rough sketch of the residual plot for these data.



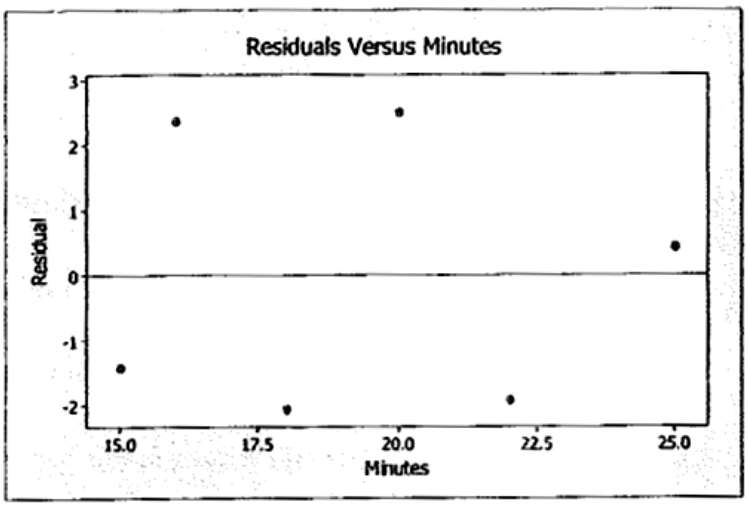


#10. Is the LSRL a good model for these data? Support your answer with appropriate evidence from your answers above.

Exercise machine - Alana's favorite exercise machine is a stair climber. On the "random" setting, it changes speeds at regular intervals, so the total number of simulated "floors" she climbs varies from session to session. She also exercises for different lengths of time each session. She decides to explore the relationship between the number of minutes she works out on the stair climber and the number of floors it tells her that she's climbed. She records minutes of climbing time and number of floors climbed for six exercise sessions. Computer output and a residual plot from a linear regression analysis of the data are shown below.

Predictor	Coef	SE Coef	T	P
Constant	-3.822	5.458	-0.70	0.522
Minutes	5.2150	0.2779	18.76	0.000

S = 2.34720 R-Sq = 98.9% R-Sq(adj) = 98.6%



#11. What is the equation of the LSRL? (Be sure to define any variables you use.)

#12. Is a line an appropriate model for these data? Justify your answer.

#13. Interpret the value of the slope of the LSRL in the context of this problem.

#14. Interpret the value of r^2 for these data in the context of this problem.

#15. What is the correlation coefficient for this data?

#16. Interpret the value of *s* (S=2.3472) in the context of this problem.

1) Given summary statistics (but no data) find the slope and LSRL

$\overline{height} = 65.8333\ in$ $\quad \overline{steps} = 16.1389$

$s_{height} = 4.9598\ in$ $\quad s_{steps} = 3.2527$ $\quad r = -.8734$

a) Use formula to calculate b.

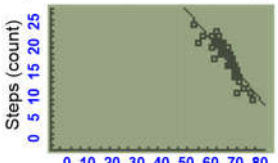
b) Find y-intercept a - the centroid $(\overline{x}, \overline{y})$ will *always* lie on the LSRL.

c) Write out the LSRL.

2) Find LSRL if x and y are swapped

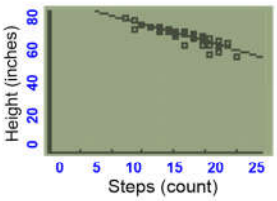
$\widehat{steps} = 53.846 - 0.5728\ height$

$\overline{steps} = 16.1389$



$r = -.8734$

$\overline{height} = 65.8333\ in$



Does r change? **no (r measures strength of association, swap does not affect r)**

Does b change? **yes** Before swap: $b = r\frac{s_y}{s_x}$ After swap: $b = r\frac{s_x}{s_y}$

3) Given data value std dev in x, find corresponding std dev in y

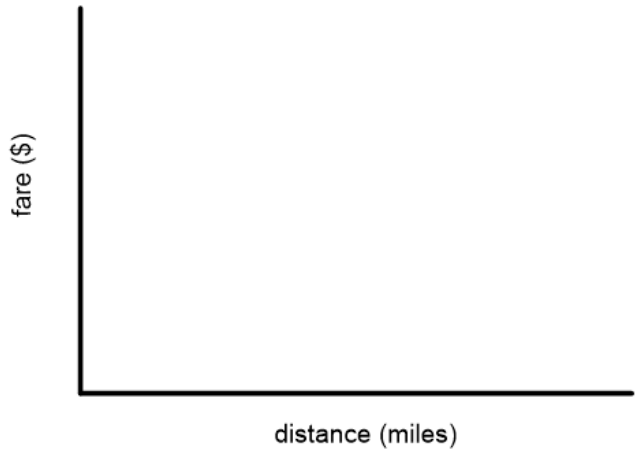
If a student's height is 1.2 standard deviations above the mean height of the class, how many standard deviations above the mean number of steps would you predict this student's number of steps to be?

(When considering standard deviations, these are z-scores, so $s_y = s_x = 1$)

Flights from Atlanta - The following data explore the relationship between the cost of an airplane ticket (fare) and the distance flown (from Atlanta to various cities):

Atlanta to:	Distance	Fare
Baltimore	568	219
Boston	933	222
Dallas	720	249
Denver	1190	308
Detroit	602	249
Kansas City	683	141
Las Vegas	1719	252
Miami	589	229
Memphis	327	183
Minneapolis	894	209
New Orleans	419	199
NY	749	248
Okla City	749	301
Orlando	392	238
Philadelphia	657	205
St Louis	461	232
Salt Lake	1565	371
Seattle	2150	343

#17. Enter the data in your calculator and sketch the scatter plot (include value labels for the axes, but do not sketch in an LSRL):



#18. Use your calculator's LinReg feature to find the LSRL:

#19. Find r^2 and r.

#20. Explain what r^2 means in this context.

#21. Explain what the slope of the LSRL means in this context.

#22. Explain what the y-intercept means in this context.

#23. Estimate the air fares for a 200-mile flight and for a 2000-mile flight.

#24. Using the estimates from #23, and #24, draw the LSRL on the scatter plot in #17.

#25. The fare to fly to Los Angeles, 1719 miles from Atlanta, is $212. Find the residual for this flight.

#26. What does a positive residual mean in this problem's context?

#27. What does a negative residual mean in this problem's context?

#28. What if we had not given you the data set, but instead only the summary statistics:

Determine the LSRL from the summary statistics (show your work).

Atlanta to:	Distance	Fare
Summary Statistics		
Mean	853.7	244.33
St Dev	497.8	56.37
Correlation	0.694	

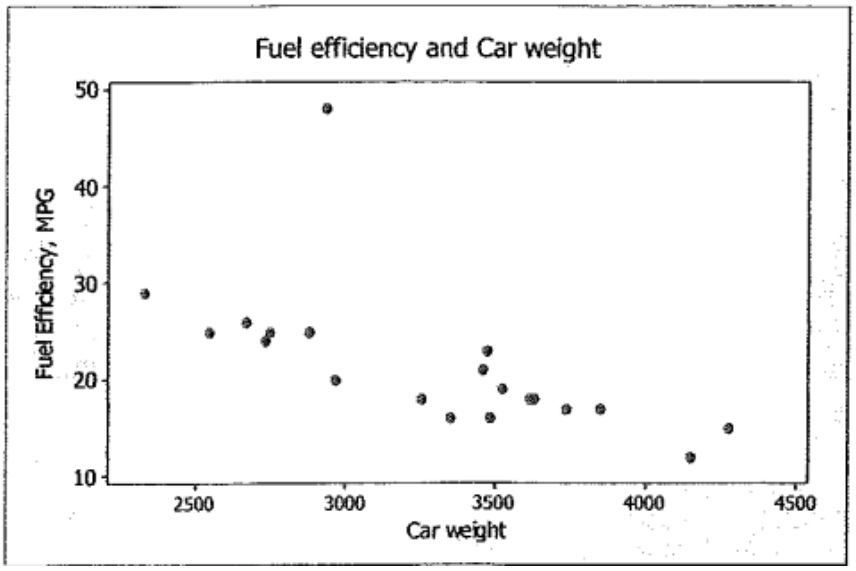
#29. What if no data or statistics were given, but you had the output of a software package?

How would you determine the LSRL and r^2 from this software output?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 177.21452   19.99315   8.864 1.43e-07 ***
distance      0.07862    0.02037   3.859  0.00139 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 41.82 on 16 degrees of freedom
Multiple R-squared:  0.482,     Adjusted R-squared:  0.4496
F-statistic: 14.89 on 1 and 16 DF,  p-value: 0.00139
```

The scatterplot below shows the fuel efficiency (in miles per gallon) and weight (in pounds) of twenty 2009 subcompact cars.



#30. There is one "unusual point" on the graph. Explain what is "unusual" about this car.

#31. What effect would removing the "unusual point" have on the slope of the LSRL for this data? Justify your answer.

#32. What effect would removing the "unusual point" have on the correlation? Justify your answer.

#33. If we converted the car weights to metric tons (1 metric ton = 2205 pounds), how would the correlation change? Explain.

Halloween Candy Sales

Year	Candy Sales (millions of dollars)
1995	1.474
1996	1.66
1997	1.708
1998	1.787
1999	1.896
2000	1.985
2001	2.035
2004	2.41

#34. How would you describe the association between the year and candy sales?

#35. What is the explanatory variable? _____

What is the response variable? _____

#36. Fit a LSRL to the data and write it down.

#37. Predict candy sales for 2003.

#38. Predict candy sales for 2013.

#39. What might cause you not to report the value you calculated in #38?

#40. What is the residual for 1998?

#41. Roughly sketch and describe the residual plot. Is a linear model appropriate for this data?

#42. What is r? _____ What does it measure?

#43. What is r^2? _____ Interpret r^2 in the context of the problem.

#44. What is the slope, b, of the LSRL? _____ Interpret b in the context of the problem.

#45. Change (1998, 1.787) to (1998, 1.387). How does this affect…

…..b? _____

……r? _____

Is this point an outlier? (Explain)

Is this an influential point? (Explain)

#46. Change (1998, 1.387) to (1998, 1.787). Add the point (2013, 0.956) How does this affect…

…..b? _____

……r? _____

Is this point an outlier? (Explain)

Is this an influential point? (Explain)

Burgers The following data were measured for a sample of burgers:

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

#47. Describe the association of calorie content vs. fat.

#48. Explanatory variable: _____

Response variable: _____

#49. Find the LSRL:

#50. Predict the calories for a burger with 30 grams of fat.

#51. Find r and r^2. Interpret r^2 in the context of the problem.

#52. Find the slope, and interpret the slope in the context of the problem.

#53. Add in a veggie burger to the data that has 36 grams of fat and 120 calories.

Is this an outlier? Is this an influential point?

What effect does adding this veggie burger have on slope?

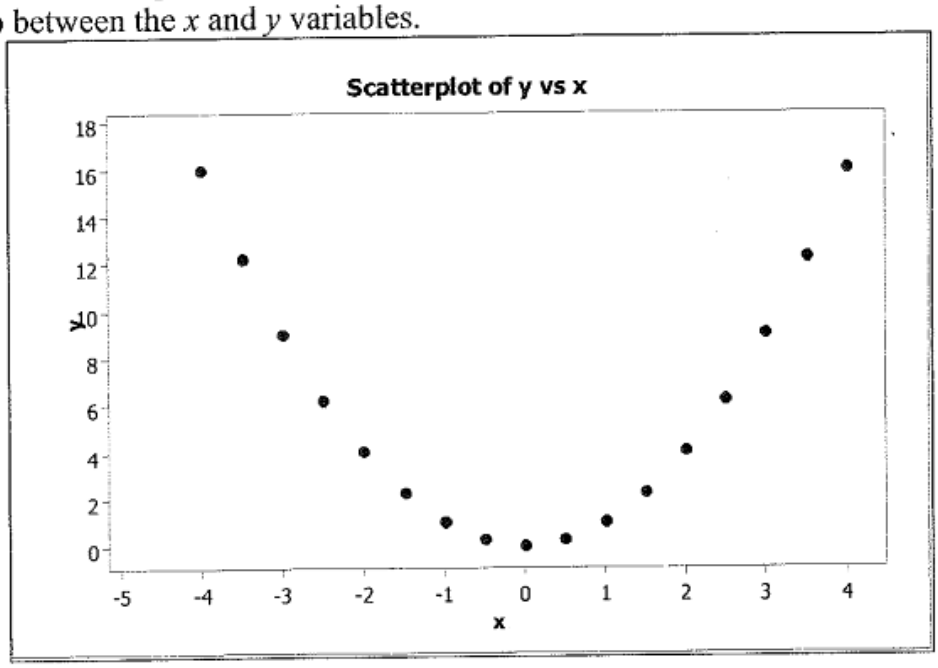
What effect does adding this veggie burger have on correlation?

What is the residual for the new data point?

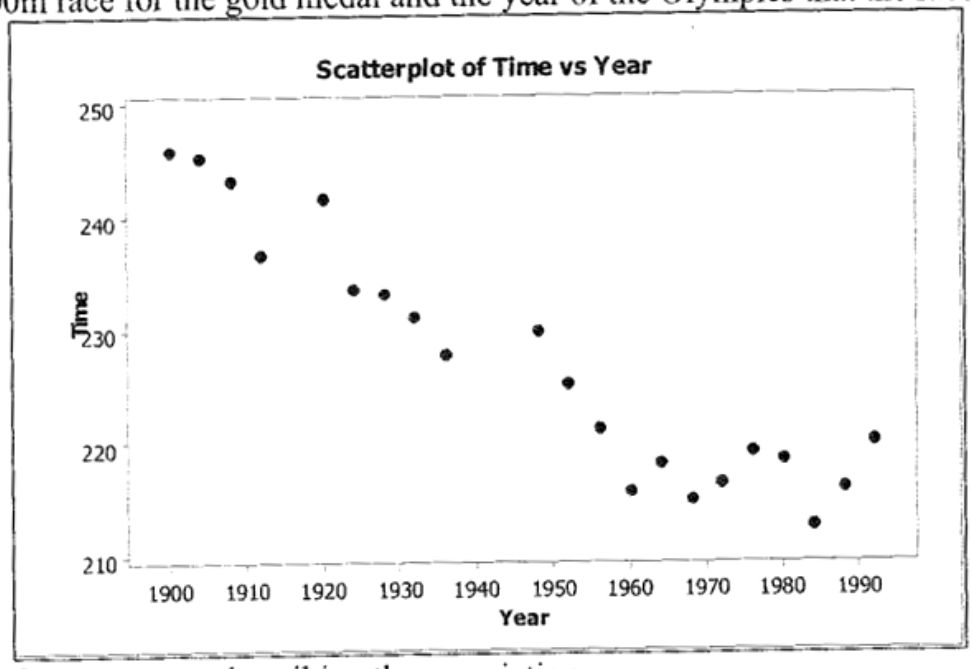
Chapter 7 Practice Quiz

1. After conducting a survey of his students, a professor reported that "There appears to be a strong correlation between grade point average and whether or not a student works." Comment on this observation.

2. The following scatterplot shows a relationship between x and y that results in a correlation coefficient of $r = 0$. Explain why $r = 0$ in this situation even though there appears to be a strong relationship between the x and y variables.



3. The following scatterplot shows the relationship between the time (in seconds) it took men to run the 1500m race for the gold medal and the year of the Olympics that the race was run in:



a. Write a few sentences describing the association.

b. Estimate the correlation. $r =$ _____

4. Identify what is wrong with each of the following statements:
   a. The correlation between Olympic gold medal times for the 800m hurdles and year is –0.66 seconds per year.

   b. The correlation between Olympic gold medal times for the 100m dash and year is -1.37.

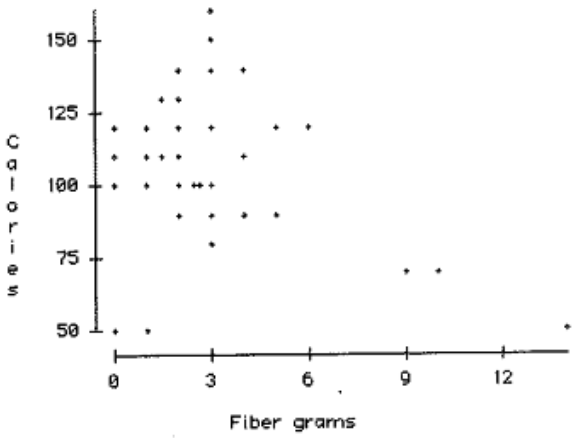
   c. Since the correlation between Olympic gold medal times for the 800m hurdles and 100m dash is –0. 41, the correlation between times for the 100m dash and the 800m hurdles is +0.41.

   d. If we were to measure Olympic gold medal times for the 800m hurdles in minutes instead of seconds, the correlation would be –0.66/60 = –0.011.

Chapter 9 Practice Quiz

Current research states that a good diet should contain 20-35 grams of dietary fiber. Research also states that each day should start with a healthy breakfast. The nutritional information for 77 breakfast cereals was reviewed to find the grams of fiber and the number of calories per serving. The scatterplot below shows the relationship between fiber and calories for the cereals.

1. Do you think there is a clear pattern? Describe the association between fiber and calories.



2. Comment on any unusual data point or points in the data set. Explain.

3. Do you think a model could accurately predict the number of calories in a serving of cereal that has 22 grams of fiber? Explain.