AP Statistics Ch 7,8,9 - Extra Practice **Name:** Solutions **Per:** ____

A student wonders if tall women tend to date taller men. She measures herself, her dormitory roommate, and the women in the adjoining rooms; then she measures the next man each woman dates. Here are the data (heights in inches):

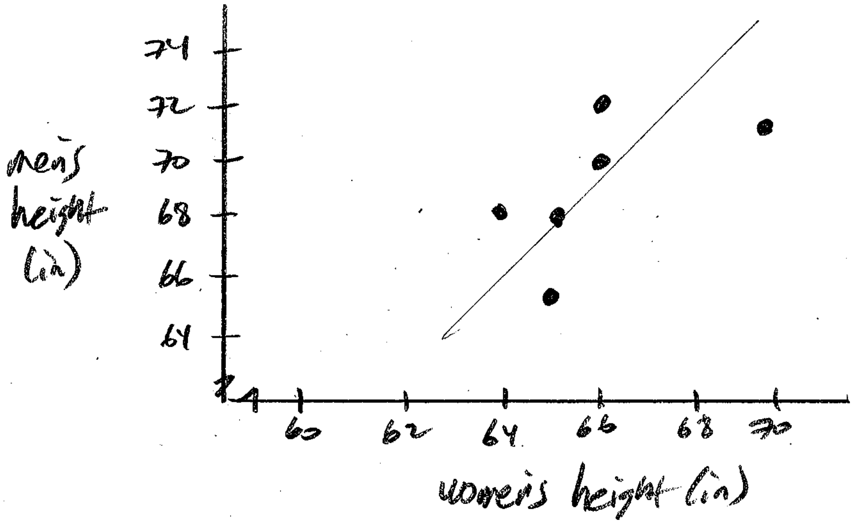
Women	66	64	66	65	70	65
Men	72	68	70	68	71	65

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

explanatory: women's height
response: men's height

"If tall women tend to date tall men" implies women are making the choice, so they are the explanatory (x) variable. (the "cause")

#2. Make a well-labeled scatterplot of these data.



#3. Based on the scatterplot, describe the pattern, if any, in the relationship between the heights of women and the heights of menthey date.

There is a fairly weak, positive, linear relationship between the heights of women and the heights of men they date.

#4. Use your calculator to find the correlation *r* between the heights of the men and women. Do the data show any evidence that taller women tend to date taller men? Explain.

$\hat{y} = 24 + .68x$ x: women's ht (in)
y: men's ht (in)

$\boxed{r = .565}$
$r^2 = .3196$

$\boxed{\text{Yes}}$, because the slope (.68 in/in) is not close to zero.
However, r=.565 indicates this is a fairly weak association. Only 32% of the variation in men's heights is explained by the LSRL model relating men's height to women's height.

#5. How would *r* change if….

…all the men were 6 inches shorter than the heights given in the table?

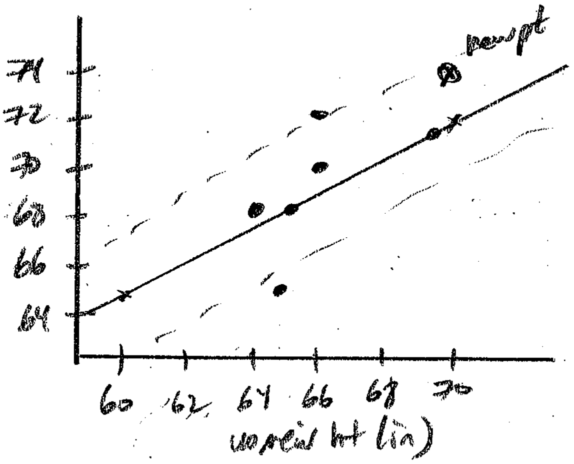
r would not change

…heights were measured in centimeters rather than inches? (There are 2.54 centimeters in each inch)

r would not change

#6. Suppose another 70-inch-tall female who dated a 73-inch-tall male were added to the data set. How would this influence *r*?

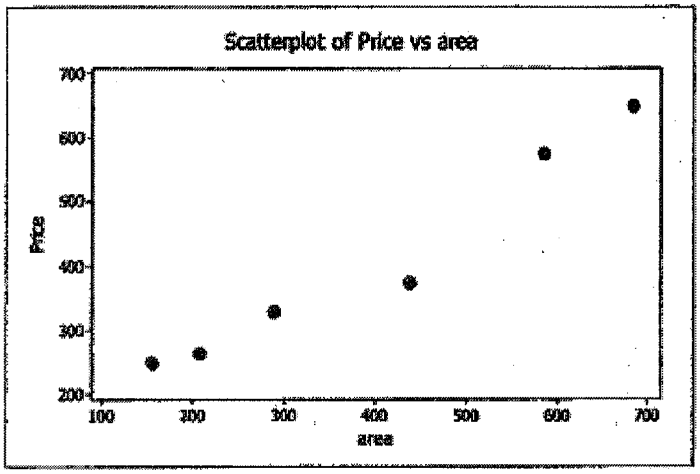
initial LSRL: $\hat{y} = 24 + .68x$
$\hat{y} = 24 + .68(60) = 64.8$
$\hat{y} = 24 + .68(70) = 71.6$
(60, 64.8) (70, 71.6) on LSRL:



This new data point is roughly the same distance as the existing data points from the LSRL so it should have little effect on r.
(adding to data, r = .565 → r = .7106)
→ so there is some effect … it is a little closer to LSRL than the most extreme points, so slightly strengthens correlation.

Below is some data on the relationship between the price of a certain manufacturer's flat-panel LCD televisions and the area of the screen. We would like to use these data to predict the price of televisions based on size.

Screen Area (sq. inches)	Price (dollars)
154	250
207	265
289	330
437	375
584	575
683	650



#7. Use your calculator to find the equation of the least-squares regression line. (define any variables used)

$\hat{y} = 105.737 + 0.769x$

x: area (in^2)

y: price ($)

#8. Explain what is meant by "least squares" in the expression "least-squares regression line".

The LSRL is the line for which the sum of the squares of the residuals is minimized.

#9. This manufacturer also produces a television with a screen size of 943 square inches. Would it be reasonable to use this equation to predict the price of that television? Explain.

No, this is extrapolation. (we can't trust that the data will remain linear)

#10. Calculate the residual for the television that has a screen area of 437 square inches. What does this number suggest about the cost of this television, relative to the others?

actual price = $375

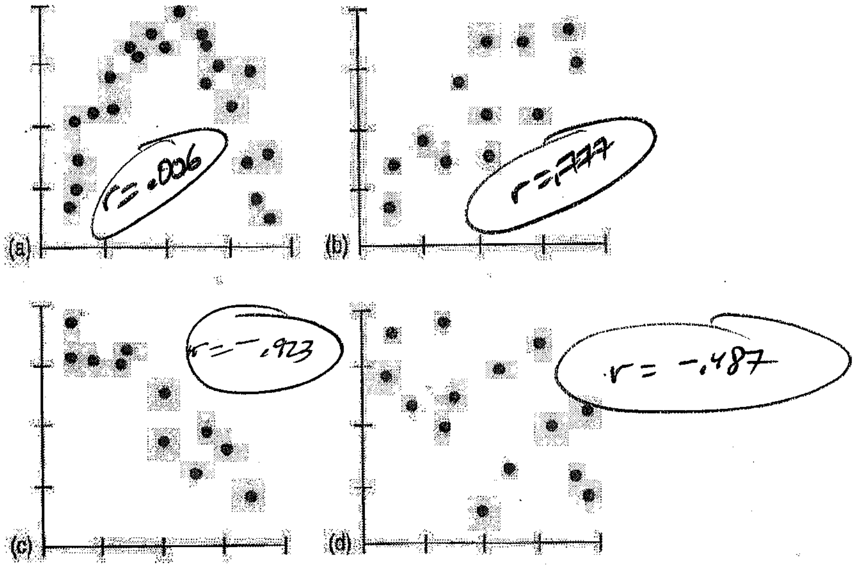
predicted $\hat{\text{price}}$ = 105.737+0.769(437) = $441.79

residual = actual − predicted = 375 − 441.79

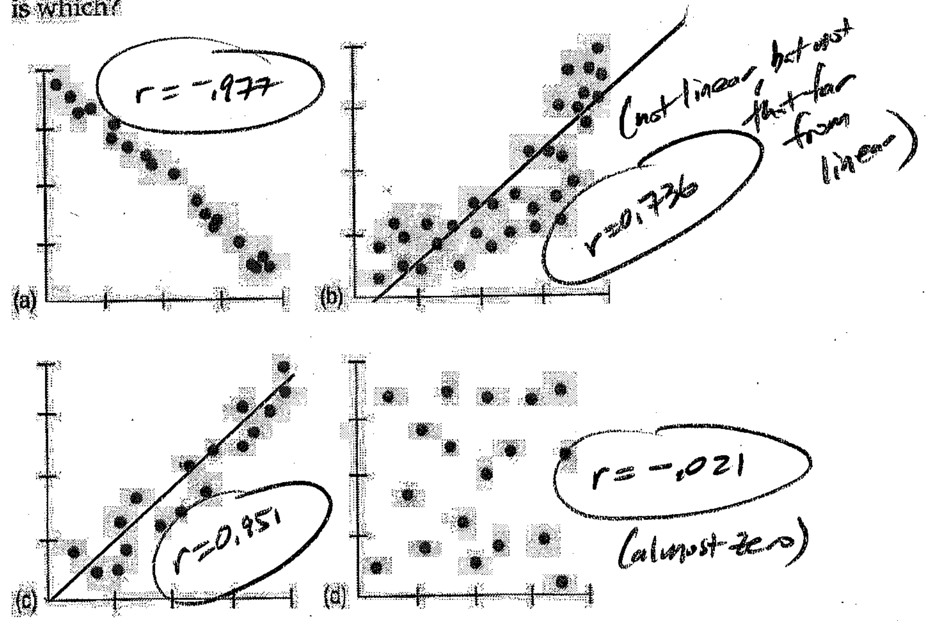
= $\boxed{-\$66.79}$

This TV is priced lower than expected for its screen size.

11. Matching. Here are several scatterplots. The calculated correlations are −0.923, −0.487, 0.006, and 0.777. Which is which?



12. Matching. Here are several scatterplots. The calculated correlations are −0.977, −0.021, 0.736, and 0.951. Which is which?



43. Body fat. It is difficult to accurately determine a person's body fat percentage without immersing him or her in water. Researchers hoping to find ways to make a good estimate immersed 20 male subjects, then measured their waists and recorded their weights.

a) Create a model to predict *%body fat* from *weight*.
b) Do you think a linear model is appropriate? Explain.
c) Interpret the slope of your model.
d) Is your model likely to make reliable estimates? Explain.
e) What is the residual for a person who weighs 190 pounds and has 21% body fat?

Waist (in.)	Weight (lb)	Body Fat (%)
32	175	6
36	181	21
38	200	15
33	159	6
39	196	22
40	192	31
41	205	32
35	173	21
38	187	25
38	188	30
33	188	10
40	240	20
36	175	22
32	168	9
44	246	38
33	160	10
41	215	27
34	159	12
34	146	10
44	219	28

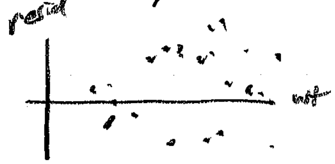
a)

% body fat
weight

$$\boxed{\begin{array}{l}\hat{y} = -27.3763+0.2499x \\ x: \text{weight (lb)} \\ y: \text{body fat (\%)}\end{array}}$$

r = .6966 r^2 = .4853

b)



$\boxed{\text{Yes}}$ because there is no pattern in the residuals.

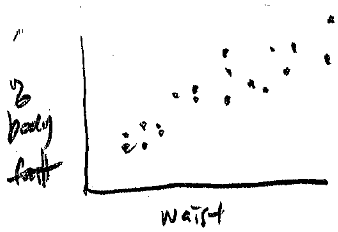
c) For every 1 additional lb in weight, % body fat increases 0.2499%, on average.

d) Yes, reasonably accurate. No outliers, r^2 = .4853, a medium-reliable model.

e) $\hat{y} = -27.3763+0.2499(190) = 20.1\%$ resid $= y - \hat{y} = 21 - 20.1 = \boxed{0.9\%}$

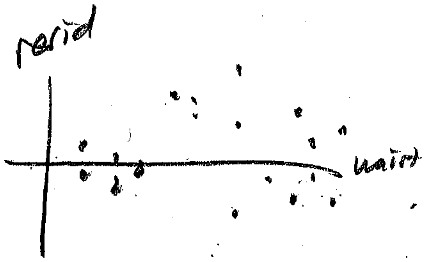
$y = 21\%$

44. Body fat, again. Would a model that uses the person's *waist* size be able to predict the *%body fat* more accurately than one that uses *weight*? Using the data in Exercise 43, create and analyze that model.



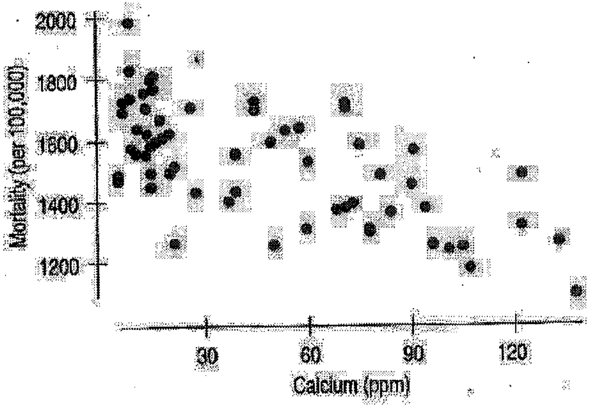
$$\boxed{\begin{array}{l}\hat{y} = -62.5573+2.2215x \\ x: \text{waist size (in)} \\ y: \text{body fat (\%)}\end{array}}$$

r = .8869 r^2 = .7865



This model is even better than using weight. The residuals plot still shows no pattern, but r^2 is higher (.7865) compared to only .4853. About 88% of the variation in % body fat is explained by the LSRL model relating % body fat to waist size.

47. Hard water. In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water. The following display shows the relationship between *mortality* and *calcium* concentration for these towns:



a) Describe what you see in this scatterplot, in context.

b) Here is the regression analysis of *mortality* and *calcium* concentration. What is the regression equation?

Dependent variable is: Mortality
R-squared = 43%
s = 143.0

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	1676	29.30	57.2	< 0.0001
Calcium	−3.23	0.48	−6.66	< 0.0001

c) Interpret the slope and *y*-intercept of the line, in context.

d) The largest residual, with a value of −348.6, is for the town of Exeter. Explain what this value means.

e) The hardness of Derby's municipal water is about 100 ppm of calcium. Use this equation to predict the mortality rate in Derby.

f) Explain the meaning of *R*-squared in this situation.

a) There is a medium strong, negative, linear association between mortality rate and calcium concentration.

b) $\hat{y} = 1676 - 3.23x$ x: calcium (ppm)
y: mortality (per 100,000)

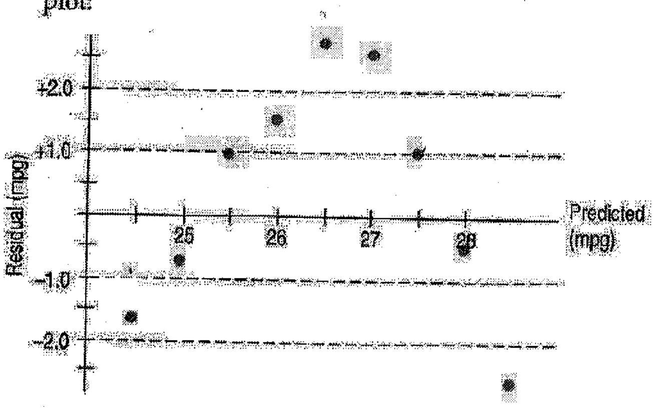
c) slope, b = −3.23
For every 1 additional ppm of calcium, there is a decrease of 3.23 deaths per 100,000 in mortality rate, on average.

d) The actual mortality rate in Exeter is 348.6 per 100,000 lower than what the model predicts the mortality rate to be.

e) $\hat{y} = 1676 - 3.23(100) = 1353$ deaths per 100,000

f) $r^2 = .43$ About 43% of the variation in mortality rate is explained by the LSRL which relates mortality rate to calcium concentration.

10. Speed. How does the speed at which you drive impact your fuel economy? To find out, researchers drove a compact car for 200 miles at speeds ranging from 35 to 75 miles per hour. From their data, they created the model $\widehat{mpg} = 32 - 0.1\ mph$ and created this residual plot:



a) Interpret the slope of this line in context.
b) Explain why it's silly to attach any meaning to the y-intercept.
c) When this model predicts high gas mileage, what can you say about those predictions?
d) What gas mileage does the model predict when the car is driven at 50 mph?
e) What was the actual gas mileage when the car was driven at 45 mph?
f) Do you think there appears to be a strong association between speed and fuel economy? Explain.
g) Do you think this is the appropriate model for that association? Explain.

a) For every 1 additional mph in speed, fuel economy decreases by 0.1 mph, on average.

b) It means for a car going 0 mph has a fuel economy of 32 mph, but this is meaningless because the car is not moving.

c) In the residuals plot, high predicted mpg is in a region where the residuals are all negative, which mean the actual mpg is lower than predicted (the model is overestimating mpg in this region).

d) $\hat{y} = 32 - 0,1(50) = \underline{27\ mpg}$

e) $\hat{y} = 32 - 0,1(45) = \underline{27.5\ mpg}$

at predicted mpg $(\hat{y}) = 27.5$, the residuals plot says the residual is +1

so $\text{resid} = y - \hat{y} \rightarrow y = 27.5 + 1 = \underline{28.5\ mpg}$

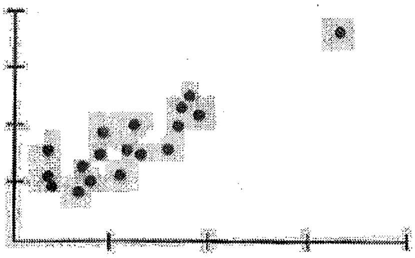
$1 = y - 27.5$

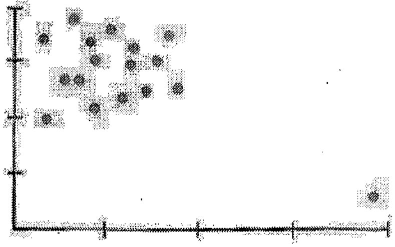
f/g) The residuals are in the range -2.5 to 2.5 mpg, so the model is predicting mpg within about 3 mpg — there must be some association.

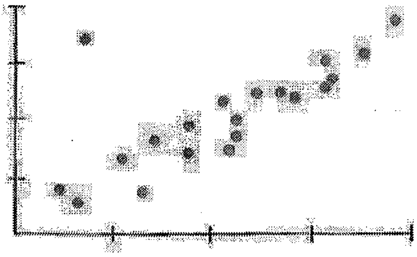
But there is a definite pattern in the residuals, so the data is <u>not</u> linear and a linear model should really be used here.

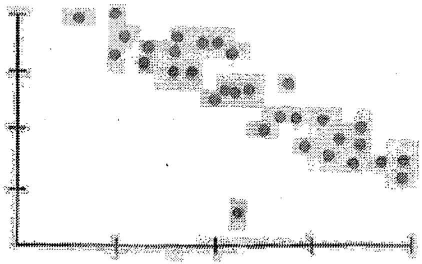
11. Unusual points. Each of the four scatterplots that follow shows a cluster of points and one "stray" point. For each, answer these questions:

1) In what way is the point unusual? Does it have high leverage, a large residual, or both?
2) Do you think that point is an influential point?
3) If that point were removed from the data, would the correlation become stronger or weaker? Explain.
4) If that point were removed from the data, would the slope of the regression line increase or decrease? Explain.


a)


b)


c)


d)

a)
- high leverage, low residual
- not influential
- r weaker (less points near the LSRL)
- b unchanged (no residual)

b)
- high leverage, probably a low residual because it would pull the LSRL to itself
- very influential
- r weaker (less points near LSRL)
- b increase (this point is the only thing making initial LSRL negative)

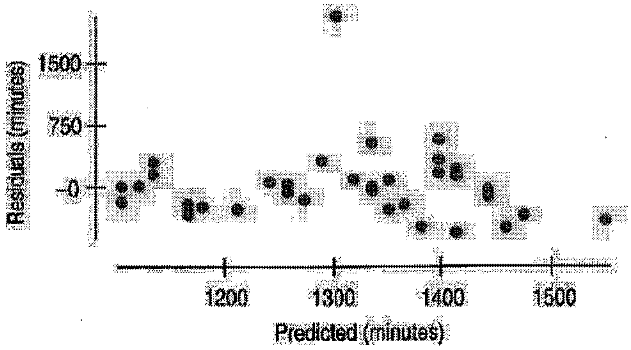
c)
- medium leverage, high residual
- moderately influential
- r strengthen (points now closer to LSRL overall)
- b slightly more positive (pt is currently pulling up left side a little)

d)
- low leverage, high residual
- not influential
- r strengthen (points now closer to LSRL overall)
- b unchanged (no leverage)

20. Swim the lake. People swam across Lake Ontario 37 times between 1974 and 2004. We might be interested in whether they are getting any faster or slower. Here are the regression of the crossing times (minutes) against the year of the crossing and the residuals plot:

Dependent variable is: Time
R squared = 7.0%

Variable	Coefficient
Intercept	−29161.9
Year	15.3323



a) What does the R^2 mean for this regression?
b) Are the swimmers getting faster or slower? Explain.
c) The outlier seen in the residuals plot is a crossing by Vicki Keith in 1987 in which she swam a round trip, North to South and then back again. Clearly, this swim doesn't belong with the others. Do you think that removing it would change the model a lot? Explain.
d) Here is the new regression after the unusual point is removed:

Dependent variable is: Time
R squared = 15.9%

Variable	Coefficient
Intercept	−28399.9
Year	14.9198

Now would you be willing to say that the swimmers were getting faster or slower?

a) About 7% of the variation in crossing times is explained by the LSRL model relating crossing time to year.

b) slope (15.33) is positive so crossing time is increasing (swimmers are getting slower) however, model is very weak $r^2 = .07$, this trend is barely noticable.

c) This point has almost no leverage. It will strengthen the correlation, but will have little effect on the slope.

d) r^2 has increased from 7% to 15.9%, so the model is somewhat stronger (but still not a very good model.)
As predicted, only small change in slope (still positive) so model still shows swimmers are getting slower.