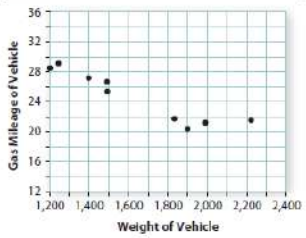
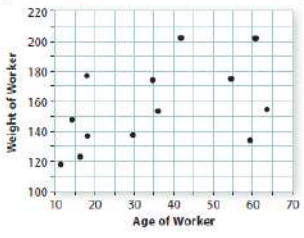
# Core II – Influential Points Worksheet (continued)

p.294/#4,5,6,CYU

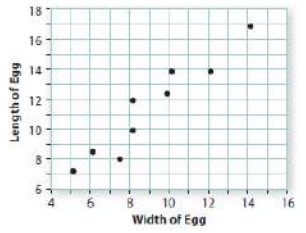
I



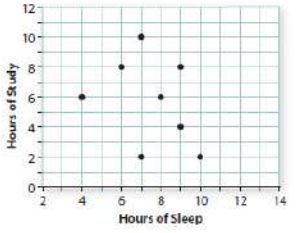

II




III




IV




4 Match each correlation with the appropriate plot. Then write a sentence that describes the association between the two variables in the plot.

**a.** $r = -0.4$

**b.** $r = 0.5$

**c.** $r = -0.8$

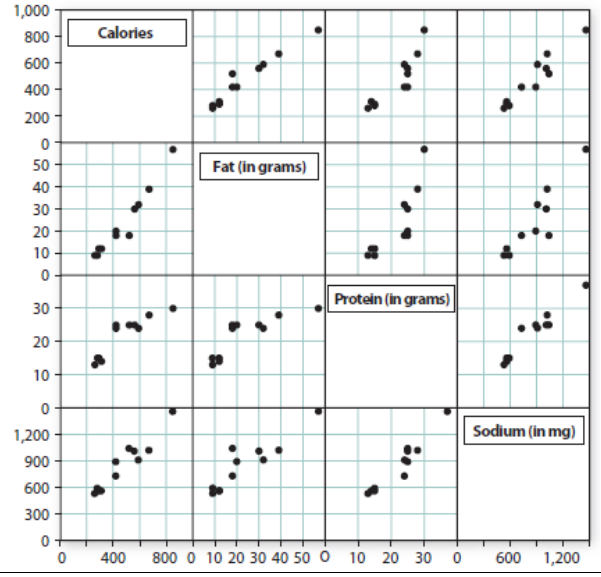
**d.** $r = 0.94$

How Hamburgers Compare

| Company | Burger | Calories | Fat (in grams) | Protein (in grams) | Sodium (in mg) |
|---|---|---|---|---|---|
| Hardee's | Hamburger | 310 | 12 | 14 | 560 |
| | Thickburger | 850 | 57 | 30 | 1,470 |
| Wendy's | Jr. Hamburger | 280 | 9 | 15 | 590 |
| | Classic Single | 420 | 20 | 25 | 880 |
| Burger King | Hamburger | 290 | 12 | 15 | 560 |
| | Whopper | 670 | 39 | 28 | 1,020 |
| McDonald's | Hamburger | 260 | 9 | 13 | 530 |
| | Quarter Pounder | 420 | 18 | 24 | 730 |
| | Big Mac | 560 | 30 | 25 | 1,010 |
| Carl's Jr. | Kid's Hamburger | 520 | 18 | 25 | 1,040 |
| | Famous Star | 590 | 32 | 24 | 910 |

**Source:** www.wendys.com; www.mcdonalds.com; www.burgerking.com; www.hardees.com; www.carlsjr.com (December 2006).

A scatterplot matrix of these data is shown below.




*Find and interpret the correlation coefficient between sodium and calories for the fast-food data. Why might this correlation be so strong?*

*Transform the amounts of sodium by converting them to grams. (1000 mg = 1 gram). Find the correlation coefficient for the transformed values of sodium and explain what you observe. Why does it make sense?*

*Now, transform the numbers of calories by subtracting 200 from each amount. Find the correlation coefficient and explain what you observe. Why does this make sense?*

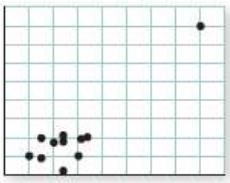
# Core II – Influential Points Worksheet (continued)

**6** As with the regression equation, you can tell if an outlier is influential on the correlation by temporarily removing it from the data set and seeing how much the correlation changes.

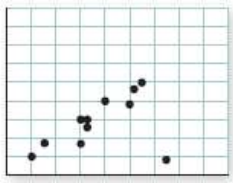
**a.** Which hamburger is an outlier in the (*sodium, calories*) data set? Is this outlier an influential point with respect to the correlation? With respect to the slope of the regression line?

**b.** For each of the plots below, identify the outlier. Indicate whether removing the point will make the correlation stronger, weaker, or unchanged.

Plot A



Plot B

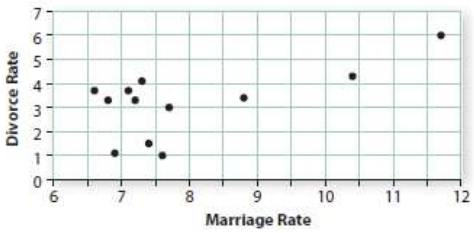


## Check Your Understanding

The table and scatterplot below give the marriage rates and divorce rates for the countries listed in the *Statistical Abstract of the United States*. Marriage and divorce rates are the number per 1,000 people aged 15–64.

| Country | Marriage Rate | Divorce Rate |
|---|---|---|
| United States | 11.7 | 6 |
| Canada | 6.8 | 3.3 |
| Japan | 8.8 | 3.4 |
| Denmark | 10.4 | 4.3 |
| France | 7.2 | 3.3 |
| Germany | 7.1 | 3.7 |
| Ireland | 7.6 | 1 |
| Italy | 6.9 | 1.1 |
| Netherlands | 7.7 | 3 |
| Spain | 7.4 | 1.5 |
| Sweden | 6.6 | 3.7 |
| United Kingdom | 7.3 | 4.1 |

**Source:** *Statistical Abstract of the United States*, 2006, Table 1320.




**a.** Estimate the correlation, and then compute the regression equation and correlation for (*marriage rate, divorce rate*).

**b.** Interpret the slope of the regression line in the context of this situation.

**c.** Which country appears to be an influential point? How will the regression equation and correlation change if this country is removed from this data set?

**d.** Remove this country, and recompute the correlation and regression equation. How influential is this country?

**e.** Convert the marriage and divorce rates for the United States to the rates per 1,000,000 people. If you do this for all countries and then recompute the regression equation and correlation, will they change?

# Core II – Influential Points Worksheet (continued)

# Core II – Influential Points Worksheet

p.287/#2,3,4

**Compact Cars**

| Car | Curb Weight (in lbs) | Highway mpg |
|---|---|---|
| Audi A4 | 3,450 | 32 |
| Chevrolet Cobalt | 3,216 | 32 |
| Ford Focus | 2,636 | 34 |
| Honda Civic | 2,690 | 40 |
| Honda Civic Hybrid | 2,875 | 51 |
| Hyundai Accent | 2,403 | 36 |
| Kia Spectra | 2,972 | 35 |
| MAZDA3 | 2,811 | 34 |
| Mercedes-Benz C280 | 3,460 | 28 |
| Nissan Sentra | 2,897 | 36 |
| Saturn ION | 2,805 | 32 |
| Subaru Impreza | 3,067 | 28 |
| Suzuki Aerio | 2,716 | 31 |
| Toyota Corolla | 2,595 | 38 |
| Toyota Yaris | 2,326 | 39 |
| Volkswagen Rabbit | 2,911 | 30 |

**Source:** www.edmunds.com

For the compact cars data, the Honda Civic Hybrid is an outlier. It has a large effect on the regression line. Remember, the original regression for this data is $y = -0.75x + 56.13$.

1) Enter the data and then delete the point for the Civic Hybrid.

*How does the regression line change?*

*How does the equation change?*

2) Replace the point for the Civic Hybrid. Now, delete the point for the Mercedes-Benz C280.

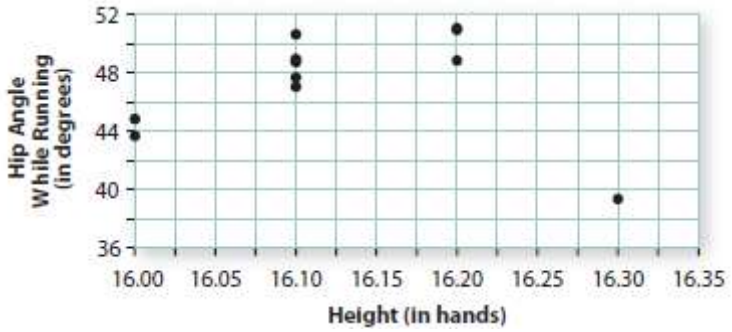
*Why are we deleting this car?*

*How does the regression line change?*

*How does the equation change?*

***Which car is more influential on the regression line and equation?***

| Horse | Height (in hands) | Hip Angle While Running (in degrees) |
|---|---|---|
| Charm | 16.1 | 47.7 |
| Hugs | 16.2 | 48.8 |
| Otis | 16.2 | 51.1 |
| Cosmo | 16.2 | 51.0 |
| Gaspe | 16.3 | 39.4 |
| Sam | 16.1 | 47.1 |
| Pi | 16.1 | 50.6 |
| Binky | 16.0 | 43.7 |
| Bella | 16.1 | 48.8 |
| Prima | 16.1 | 48.7 |
| Bandit | 16.0 | 44.8 |
| Blackie | 16.1 | 48.9 |




*Identify the horse that is an outlier:*

Remove the horse you identified from the data set.

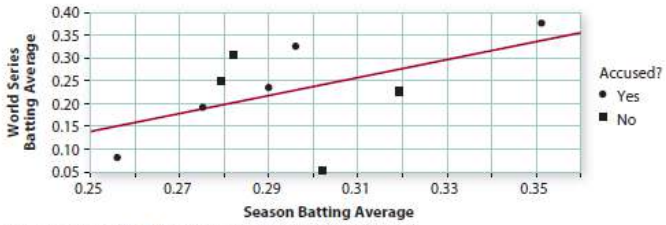
*What happened to the slope and intercept of the regression line?*

*Can you say that this horse is influential? Why?*

# Core II – Influential Points Worksheet

Chicago White Sox

| Player | Season Batting Average | World Series Batting Average | Accused? N=no/Y=yes |
|---|---|---|---|
| Eddie Collins | .319 | .226 | N |
| Shano Collins | .279 | .250 | N |
| Happy Felsch | .275 | .192 | Y |
| Chick Gandil | .290 | .233 | Y |
| Shoeless Joe Jackson | .351 | .375 | Y |
| Nemo Leibold | .302 | .056 | N |
| Swede Risberg | .256 | .080 | Y |
| Ray Schalk | .282 | .304 | N |
| Buck Weaver | .296 | .324 | Y |




Source: www.baseball-reference.com/postseason/1919_WS.shtml

This data shows the 1919 season and World Series batting averages for the nine White Sox players who had 10 or more at bats in the World Series. Five of these players are accused of throwing the series to the Cincinnati Reds.

The regression line is $y = 1.99x - 0.36$.

*Which player did the worst in the World Series, compared to what would be predicted? Was he one of the accused?*

*Find out if the players below are influential. For each player, remove their data point and explain how the regression line and equation changed.*
*(**remember to replace the previous player before removing the next player!** )*

*Nemo Leibold:* _____

*Swede Risberg:* _____

*Shoeless Joe:* _____

**http://www.wmich.edu/cpmp/CPMP-Tools/**

# Unit 4 Review – Summary Notes & Practice

In this unit, you began by looking at scatterplots.

- Explain why you should always look at a scatterplot before computing a regression line or interpreting a correlation.

- What does it mean when a scatterplot is called "linear"?

The summary line that you computed is called the least squares regression line.

- Give two reasons that you might want to find a regression line for a set of paired data.

- What is the meaning of "least squares"?

- Explain what a residual is and how it can be estimated from the scatterplot.

- Explain how a residual can be found from the regression equation.

The correlation is another summary statistic for paired data.

- What does computing a correlation add to the summary provided by the least squares regression line?

- Explain how you can compute the correlation for a set of paired data.

It is always important to consider influential points when interpreting regression and correlation.

- How do you decide if a point is influential?

Sketch a scatterplot that illustrates each simple description below. Give a pair of variables that might match each scatterplot.

strong positive correlation

strong negative correlation

weak positive correlation

weak negative correlation

Sketch a scatterplot with an influential point that:

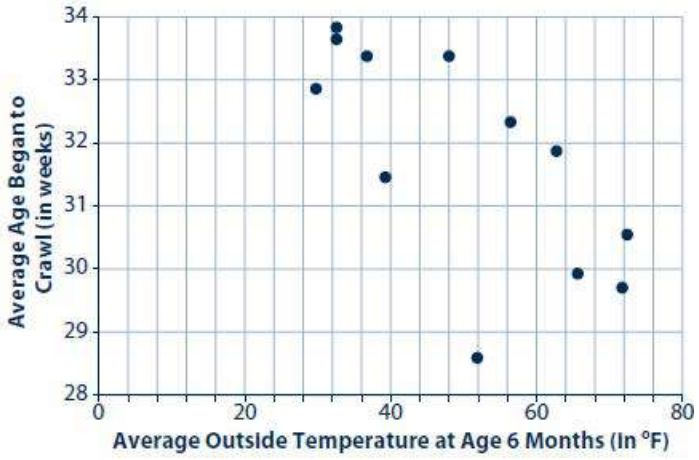
increases the correlation

decreases the correlation

# Unit 4 Review – Summary Notes & Practice

p.323/#1

| Birth Month | Average Outside Temperature at Age 6 Months (in °F) | Age Began to Crawl (in weeks) |
|---|---|---|
| January | 66 | 29.84 |
| February | 73 | 30.52 |
| March | 72 | 29.70 |
| April | 63 | 31.84 |
| May | 52 | 28.58 |
| June | 39 | 31.44 |
| July | 33 | 33.64 |
| August | 30 | 32.82 |
| September | 33 | 33.83 |
| October | 37 | 33.35 |
| November | 48 | 33.38 |
| December | 57 | 32.32 |




**a.** Does it appear from the scatterplot that babies who are six months old during cold months of the year learn to crawl at a later age on average than babies who are six months old during warmer months?

**b.** Approximately how many babies are represented by each point on the scatterplot?

**c.** What is the shape of the cloud of points?

**d.** Find the least squares regression line for predicting age from temperature, and graph it on a copy of the scatterplot.

**e.** Interpret the slope of the regression line in the context of these data.

**f.** Find the point that has the largest residual (in absolute value).

- **i.** In what month were these babies born?
- **ii.** Estimate the residual for that point from the scatterplot.
- **iii.** Compute this residual using the regression equation and the data in the table.
- **iv.** Is this point an outlier in terms of $x$, in terms of $y$, in terms of both $x$ and $y$, or only in terms of $x$ and $y$ jointly? Explain.

**g.** Is the point you identified in Part f an influential point? Explain your reasoning.

# Unit 4 Review – Summary Notes & Practice