ANALYZING RELATIONSHIPS: MARRIAGE, DIVORCE, AND LINEAR REGRESSION

TEACHER VERSION

Subject Level:
High School Math

Grade Level:
9

Approx. Time Required:
180 minutes

Learning Objectives:

- Students will be able to assess how well a linear model fits the data by plotting and analyzing the residuals.
- Students will be able to determine the impact of outliers on the linear model.
- Students will be able to explain the meaning of the slope and y-intercept of the linear model in the context of the data.

Activity Description

Students will examine census data on marriage and divorce rates for women and men in each state and the District of Columbia. From these data, they will create a scatter plot, find a line of best fit, and analyze the relationship between the two variables (i.e., sex and marriage/divorce rates). They will also use a residual plot, explain the meaning of the slope and of the y-intercept of the line of best fit, and investigate the effect of outliers on this line.

Suggested Grade Level:
9

Approximate Time Required:
180 minutes

Learning Objectives:

- Students will be able to assess how well a linear model fits the data by plotting and analyzing the residuals.
- Students will be able to determine the impact of outliers on the linear model.
- Students will be able to explain the meaning of the slope and y-intercept of the linear model in the context of the data.

Topics:

- Linear model
- Linear regression
- Residuals and residual plots
- Scatter plots
- Slope
- Y-intercept

Skills Taught:

- Analyzing a scatter plot and a residual plot
- Interpreting slope and y-intercept
- Making predictions using a linear model

Materials Required

- The student version of this activity, 13 pages
- Graphing calculators

Activity Item

The following item is part of this activity. The item, its data source, and instructions for viewing the source data online appear at the end of this teacher version.

- Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia

For more information to help you introduce your students to the U.S. Census Bureau, read "*Census Bureau 101 for Students*." This information sheet can be printed and passed out to your students as well.

Standards Addressed

See charts below. For more information, read "*Overview of Education Standards and Guidelines Addressed in Statistics in Schools Activities*."

Common Core State Standards for Mathematics

Standard	Domain	Cluster
CCSS.MATH.CONTENT.HSS.ID.B.6 Represent data on two quantitative variables on a scatter plot, and describe how the variables are related.	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on two categorical and quantitative variables.
CCSS.MATH.CONTENT.HSS.ID.B.6.A Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models.	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on two categorical and quantitative variables.
CCSS.MATH.CONTENT.HSS.ID.B.6.B Informally assess the fit of a function by plotting and analyzing residuals.	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on two categorical and quantitative variables.
CCSS.MATH.CONTENT.HSS.ID.B.6.C Fit a linear function for a scatter plot that suggests a linear association.	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on two categorical and quantitative variables.
CCSS.MATH.CONTENT.HSS.ID.C.7 Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data.	ID – Interpreting Categorical & Quantitative Data	Interpret linear models.

Common Core State Standards for Mathematical Practice

Standard
CCSS.MATH.PRACTICE.MP4. Model with mathematics. Students will fit a linear model to a set of data, explaining the meaning of the parameters of the model in the context of the data, assessing how well the model fits the data, determining the impact of an outlier on the model, and using the model to make predictions.
CCSS.MATH.PRACTICE.MP6. Attend to precision. Students will communicate precisely about the data, especially when explaining the meaning of the slope of a linear model.

National Council of Teachers of Mathematics' Principles and Standards for School Mathematics

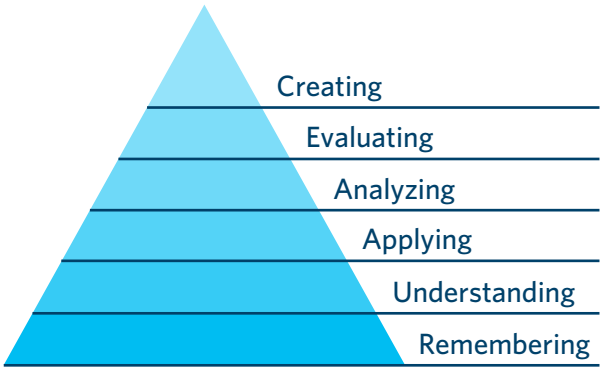
Content Standard	Students should be able to:	Expectation for Grade Band
Algebra	Use mathematical models to represent and understand quantitative relationships.	Draw reasonable conclusions about a situation being modeled.
Data Analysis and Probability	Select and use appropriate statistical methods to analyze data.	For bivariate measurement data, be able to display a scatterplot, describe its shape, and determine regression coefficients, regression equations, and correlation coefficients using technological tools.
Data Analysis and Probability	Select and use appropriate statistical methods to analyze data.	Identify trends in bivariate data and find functions that model the data or transform the data so that they can be modeled.

Guidelines for Assessment and Instruction in Statistics Education

GAISE	Level A	Level B	Level C
Formulate Questions		X	
Collect Data			
Analyze Data	X		
Interpret Results	X		

Bloom's Taxonomy

Students will ***analyze*** data in a linear model, ***apply*** the model to make predictions, and ***evaluate*** the effectiveness of the model.



Teacher Notes

Before the Activity

Students must understand the following key terms:

- **Association** – a relationship between two variables that can be weak or strong, positive or negative, or nonexistent; the form of an association can also be linear or nonlinear.
- **Linear model** – a line of best fit that is used to predict *y* values based on *x* values
- **Linear regression** – an approach for modeling the relationship between *y* and *x* values
- **Outlier** – an extremely high or low value that noticeably differs from the other data points in the set
- **Residual** – the difference between the actual *y* coordinate of a data point and what the linear model predicts (actual - predicted)
- **Residual plot** – a scatter plot of all the residuals
- **Slope** – the rate of change in a linear model or the strength and direction of association between variables in a linear regression

Students should have the following skills:

- Ability to make a scatter plot using graphing technology
- Ability to find the line of best fit
- Ability to find predicted values using a line of best fit

Students must understand the following idea:

- Scatter plots can contain unusual observations that may or may not influence the relationship between the variables.

Teachers should explain to students that data from this activity come from the American Community Survey, which is conducted monthly by the Census Bureau and is designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment.

Teachers should also explain to students that, once they start the activity, they will analyze data on marriage and divorce rates for women and men in each U.S. state and the District of Columbia. Teachers should divide students into groups of two to four, asking them to work together to make predictions about how they think marriage and divorce rates in the United States have changed between 2008 and 2014. Teachers should ask student groups to share their predictions, recording them on the board, on chart paper, or somewhere else visible in the classroom. Teachers should ask these questions to get students thinking:

- Do you think the rates will be high or low overall?
- Do you think they will be about the same for all states and the District of Columbia? What about for women and men?

- Which states and sexes do you think will have higher or lower rates for marriage? For divorce? Are there any states whose rates you think will be extremely high or low?

Teachers should then ask students to talk with their groups about why they think the data include rates instead of actual numbers of marriages and divorces. Then teachers should ask student groups to share their thinking. Teachers should look for student answers mentioning that rates allow for comparison across states with different population sizes — more populous states will generally have more marriages and divorces than less populous states, but the rates of marriage and divorce could be similar.

During the Activity

In part 1, teachers can either have students help them come up with a mathematical question to investigate or provide students with the teacher-written question offered below.

If teachers would like students' help in coming up with a question, they should follow these steps.

1. Ask them to individually examine **Item 1** and write down one observation and one question. Look for observations like: The 2014 marriage rates tend to be lower than the 2008 marriage rates; the 2008 divorce rates tend to be higher than the 2014 divorce rates. Look for questions like: Do the 2008 marriage rates predict the 2014 rates? Why are the marriage rates for men higher than the marriage rates for women, while the divorce rates for men are lower than the divorce rates for women?
2. Ask students to share their observations and questions, and record them on the board. Then ask them to identify which questions they might be able to answer by examining the association between the two variables in the data.
3. Choose one question for the whole class to investigate in the activity. Questions could include: How are the 2008 marriage rates for men and women associated? How are the 2014 divorce rates for men and women associated?

If teachers would like to use the mathematical question and sample answers provided in the activity, they should follow these steps.

1. Have students examine the 2014 marriage data for men and women in **Item 1**.
2. Ask students these questions and discuss: Do states with lower marriage rates for men also tend to have lower marriage rates for women, and vice versa? How easy is it to predict marriage rates for women by looking at marriage rates for men? Which states tend not to follow the general pattern you see when looking at the data?
3. Tell students to write down the following question as the answer to the first prompt under part 1: How are the 2014 marriage rates for men and women associated?

Teachers will then direct students to complete the activity in groups, monitoring them as they work.

After the Activity

Teachers should ask the following questions to facilitate a class discussion reinforcing key lessons learned:

- How can you tell if a linear model is appropriate for the association between two variables in a scatter plot? (Teachers should expect answers like: If the cloud of points formed by the scatter plot is roughly oval-shaped with an overall constant trend, then a linear model is appropriate — the thinner the oval, the stronger the linear relationship.)
- How can you tell if a linear model is appropriate by looking at the residual plot? (Teachers should expect answers like: You know a linear model is appropriate when the residual plot shows no particular pattern and is evenly scattered above and below the horizontal line at 0.)
- How do you explain the meaning of the slope of a linear regression as compared with that of a regular linear function? (Teachers should expect answers like: The slope of a linear regression represents a comparison — not the average rate of change, as in a regular linear function.)

Extension Idea

- Teachers could have students compare current data on marriage and divorce rates for men and women in a particular state — perhaps the state where they live — using ***data.census.gov***.

Student Activity

Click ***here*** to download a printable version for students.

Student Learning Objectives

- I will be able to assess how well a linear model fits the data by plotting and analyzing the residuals.
- I will be able to determine the impact of outliers on the linear model.
- I will be able to explain the meaning of the slope and of the y-intercept of the linear model in the context of the data.

Activity Item

The following item is part of this activity and appears at the end of this student version.

- Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia

Part 1 – Explore Data in Scatter and Residual Plots

1. Write down the question your class will answer by examining census data.

   How are the 2014 marriage rates for men and women associated?

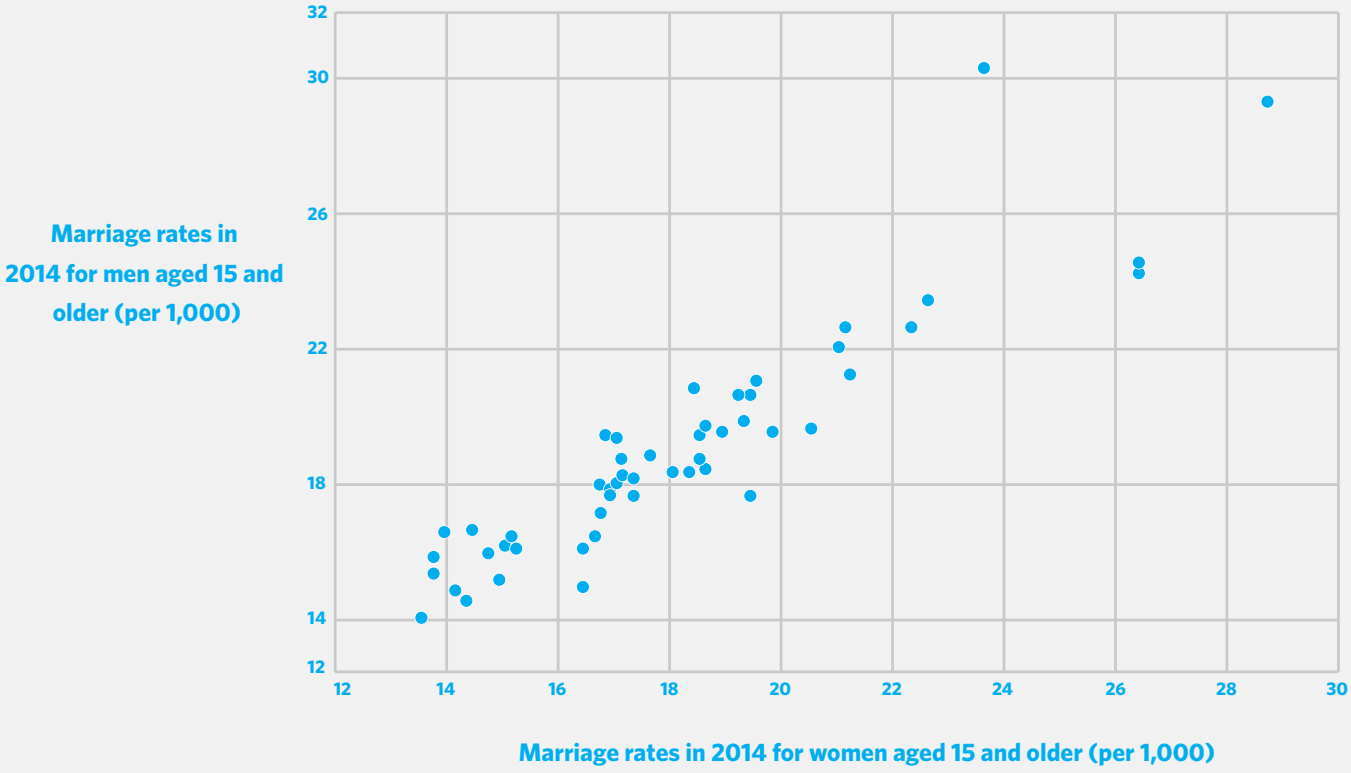
2. Which data from **Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of the Columbia** will help you answer your question?

   Marriage rates for women in 2014 and marriage rates for men in 2014 (for all 50 U.S. states and the District of Columbia).

3. Make a scatter plot of your data on the grid below.

   See sample scatter plot below.

Comparing 2014 Marriage Rates for Women and Men in the 50 U.S. States and the District of Columbia



4. Consider which of your variables you plotted on the horizontal and vertical axes. Did your choice matter in this situation? Why or why not?

   No. There is no independent-dependent relationship between these two variables.

5. Choose any data point on your scatter plot to examine. What does that point represent in the context of the data set?

   Student answers will vary but could include: The point I chose (17.0, 19.4) represents the marriage rates for women and men, respectively, in Alabama. This point shows that, on average, 17.0 out of 1,000 women and 19.4 out of 1,000 men were married in Alabama in 2014.

   a. Do the two variables appear to have an association?

      Yes.

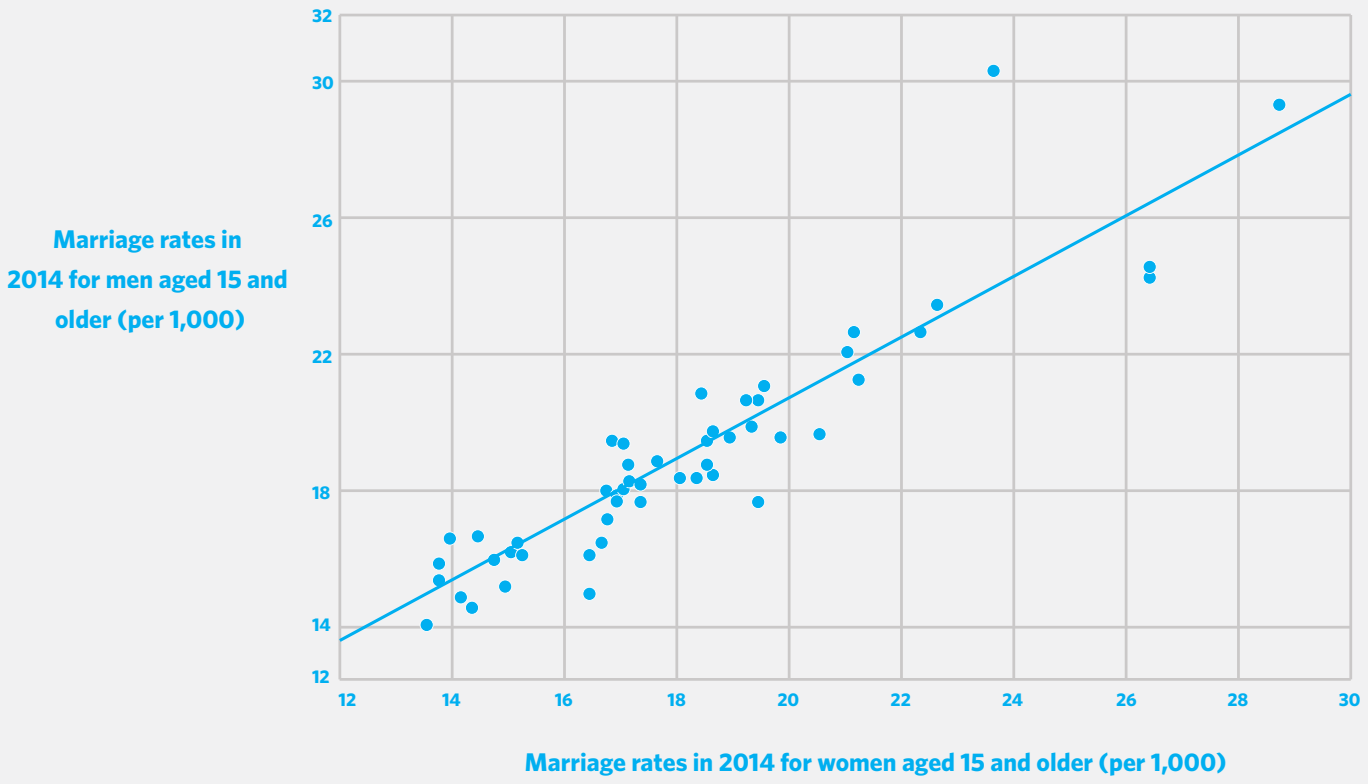
b. If so, is it linear? Weak or strong? Positive or negative? How do you know?

It is a strong, positive linear association because the points are fairly close to a linear pattern and the *y* values increase as the *x* values increase.

6. Add a line of best fit to your scatter plot and use your graphing calculator to determine the equation for that line. Keep in mind that statisticians typically replace the generic variables *x* and *y* with the actual variable names, or at least descriptive variable names. They also place a " ^ " symbol over the *y* variable to indicate that it is a predicted value. For example, an equation relating 2010 marriage rates for women (as the *y* variable) and for men (as the *x* variable) might be: $\widehat{\text{women}} = 0.94 \cdot \text{men} + 0.08$. Write your equation below.

$$\widehat{\text{men}} = 0.92 \cdot \text{women} + 2.33$$

Comparing 2014 Marriage Rates for Women and Men in the 50 U.S. States and the District of Columbia



7. A residual is the difference between the actual *y* coordinate of a data point and what the linear model predicts (actual - predicted). Calculate the residual for any two points on your scatter plot, using your line of best fit to determine predicted values. Show your work.

   Student answers will vary but could include the following:

   - Point 1: **Indiana: (19.3, 19.9); $\widehat{men} = 0.92 \cdot 19.3 + 2.33 = 20.0$. The predicted value for men is 20.0, and the actual value is 19.9. So the residual (actual - predicted) is -0.1 marriages per 1,000 men.**
   - Point 2: **Washington, D.C.: (23.6, 30.4); $\widehat{men} = 0.92 \cdot 23.6 + 2.33 = 24.0$. The predicted value for men is 24.0, and the actual value is 30.4. So the residual (actual - predicted) is 6.4 marriages per 1,000 men.**

8. If the residual is positive, what does that mean and how does it compare with the line of best fit? What about if the residual is negative?

   If the residual is positive, then the actual value is larger than the predicted value and the point appears above the line of best fit. If the residual is negative, then the actual value is smaller than the predicted value (overpredicted) and the point appears below the line of best fit.

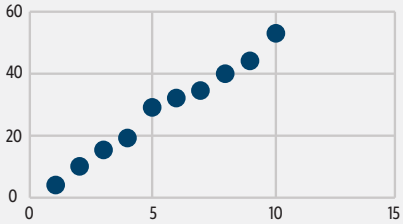
9. What does it mean if the residual is close to 0? What if the residual isn't close to 0? How would each of these residuals appear in a scatter plot?

   If the residual is close to 0, this means the actual value is very close to the predicted value and the point appears close to the line of best fit. If the residual isn't close to 0, it could be an outlier (larger residual) and appear farther away vertically from the line of best fit (or further off the general trend of the rest of the points).

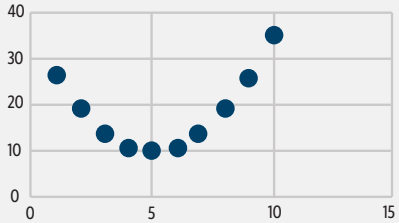
10. You can determine whether a linear model is appropriate for your data by looking at the residual plot, which is a scatter plot of all the residuals versus the *x* variable. The residual plot shows the relationship between the actual data after accounting for the linear pattern. If a linear model is appropriate for the data, there should be no remaining pattern in a residual plot and the points should be spread equally above and below the horizontal line at 0.

    With this information in mind, draw a line to connect each of the three scatter plots with what you think is its residual plot below.

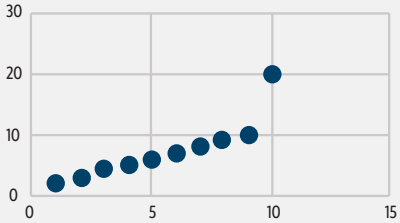
Scatter Plots



A

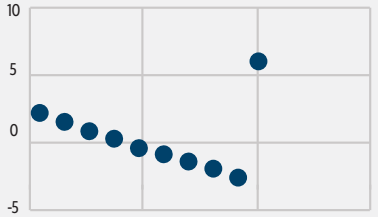


B

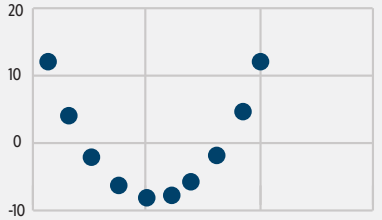


C

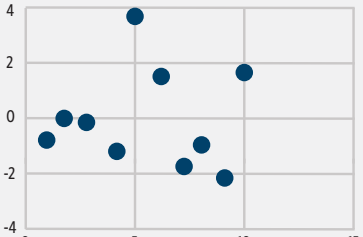
Residual Plots



A



B



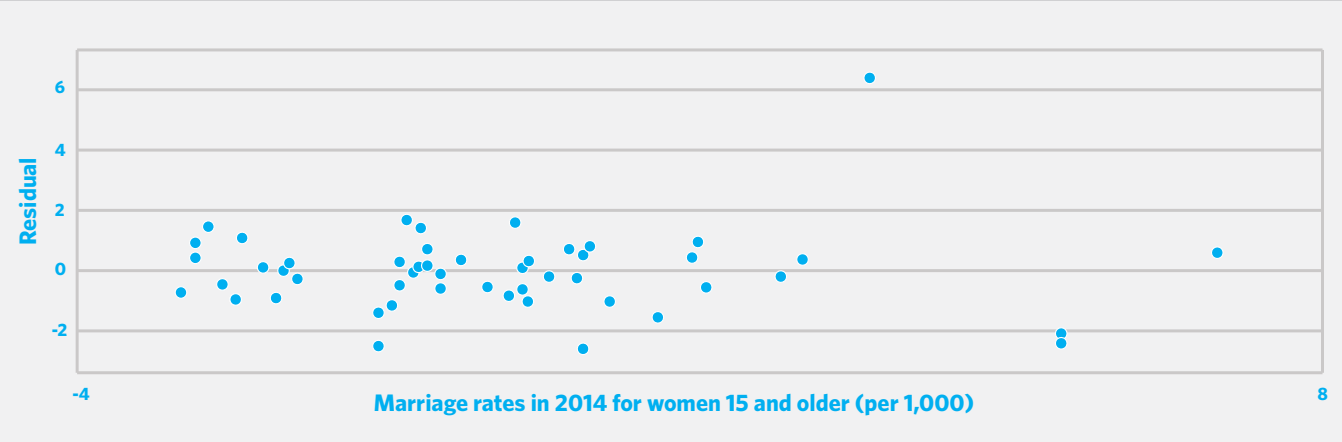
C

- **Scatter plot A goes with residual plot C.**
- **Scatter plot B goes with residual plot B.**
- **Scatter plot C goes with residual plot A.**

  Teachers should check in with students to make sure they understand why residual plot C represents a good linear fit, while residual plots A and B do not.

11. Using the data from **Item 1** and your line of best fit from question 6, make a residual plot in the grid below. Draw a horizontal line at 0, and graph the residual for each data point above or below this line, depending on whether it is positive or negative.

Comparing 2014 Marriage Rates for Women and Men in the 50 U.S. States and the District of Columbia



12. Is a linear model appropriate for your data? Explain.

Yes. The residual plot shows no distinct pattern (apart from a few outliers).

13. You already know that the slope of a linear model represents the rate of change of the line.

Statisticians talk about slope in slightly different terms: as a measure of comparison between two *x* values. In the 2010 example equation from question 6 ($\widehat{women}$ = 0.94 • men + 0.08) the slope is 0.94. This means that if state A has 10 more marriages per 1,000 men than state B, then state A is likely to have approximately 9.4 more marriages per 1,000 women (i.e., 10*0.94) than state B.

What is the slope of your line of best fit? Explain what the slope means in the context of comparing two points in your data set.

0.92. Student answers should refer to slope as a measure of comparison between two states (the predicted difference in their *y* values based on their *x* values).

14. You also already know that the y-intercept represents the predicted *y* value when the *x* value equals 0. Use that knowledge to answer the following questions and prompts.

a. What is the y-intercept for your line of best fit, rounded to the nearest hundredth?

The y-intercept is 2.33 marriages per 1,000 men.

b. Explain what the y-intercept means in the context of your data.

It represents the predicted marriage rate for every 1,000 men when the marriage rate for every 1,000 women is 0.

c. Does it make sense to interpret the y-intercept in this way? Explain.

No. In the context of this data, it does not make sense for the marriage rate for men to be a value other than 0 when the marriage rate for women is 0.

15. Some scatter plots contain outliers, which are data points that fall outside of the pattern of the rest of the points (i.e., far from the line of best fit). Looking at both the scatter plot and the residual plot you created, can you identify any outliers? Be sure to explain how you used your residual plot to confirm any outliers.

There appears to be one outlier (23.6, 30.4) well above the line and separate from all other residuals. A residual plot helps to identify outliers by showing any point that has a very large (positive or negative) residual compared with other residuals. The point appears to be an outlier in the scatter plot, and the residual plot confirms this by showing that the point has a residual of more than 6, while most other residuals are clustered within 2 units of 0.

16. If your scatter plot contains an outlier, identify which U.S. state (or the District of Columbia) that point represents. (If your scatter plot has multiple outliers, pick the greatest one.) Why do you think this location displays this unusual behavior in the context of the data? If your scatter plot does not contain an outlier, explain how you know that.

The outlier represents the District of Columbia. Student answers will vary for why the location might display unusual behavior in the context of the data.

17. Sometimes an outlier can influence the linear model by pulling the line of best fit toward it. In some cases, outliers make the association appear stronger; in other cases, it appears weaker. One way to determine whether an outlier is influencing the linear model is to remove it from the data set and see what happens.

    If your scatter plot has any outliers, remove these data points from the set and recalculate the new line of best fit. Write it below. Does the removal of outliers change the line significantly? Why or why not?

The new equation is: $\widehat{men} = 0.85 \cdot women + 3.42$. The removal of the outlier doesn't change the line significantly because the point was already very close to the line of best fit and had a small residual.

18. How do the components of your new line of best fit and its equation compare with those of the original line and equation?

The slope of the new line is slightly flatter than the slope of the original, while the y-intercept is a little larger than the original.

19. Put the point you removed in question 17 back in the data set and select another point, close to the middle of your data set, to remove. Recalculate the equation for this line of best fit and write it below. How does this new line compare with the original? How does it compare with the line you calculated for question 17, and why do you think that is?

    Student answers will vary but could include: When I removed another point (21.0, 22.1), which represents Colorado, the equation became: $\widehat{men}$ = 0.92 * women + 2.37. This new line is very similar to the original line. This new line is different from the line I calculated for question 17, which indicates that the removal of the outlier in question 17 had more of an influence on the data than the removal of a more central data point.

Part 2 – Make Predictions

Linear models can be used to make predictions. With that in mind, predict the 2010 marriage rate for women in Puerto Rico using the example equation from earlier ($\widehat{women}$ = 0.94 • men + 0.08) and the following information: The 2010 marriage rate for men in Puerto Rico is 10.7 marriages per 1,000 men.

1. Write the new equation that you will use to make your prediction. What value does the linear model predict?

   $\widehat{women}$ = 0.94 • 10.7 + 0.08 = 10.1. The linear model predicts that there were 10.1 marriages per 1,000 women in Puerto Rico in 2010.

2. The actual 2010 marriage rate for women in Puerto Rico is 9.2 marriages per 1,000 women. How close was your prediction in question 1 to the actual value? How confident are you in this prediction?

   The prediction was close, with a residual of -0.9 marriages per 1,000 women (i.e., 9.2 - 10.1). Because of this, I am fairly confident in my prediction.

Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia

State	Marriage Rates				Divorce Rates			
	2014		2008		2014		2008	
	Women	Men	Women	Men	Women	Men	Women	Men
Alabama	17.0	19.4	20.7	22.5	10.4	9.4	13.5	13.7
Alaska	28.7	29.4	24.8	27.8	10.3	9.1	12.5	13.2
Arizona	17.3	17.7	16.8	18.6	8.8	7.5	12.0	11.3
Arkansas	20.5	19.7	24.5	25.6	11.6	12.0	14.1	11.1
California	17.1	18.3	17.8	19.5	7.2	6.7	9.9	8.3
Colorado	21.0	22.1	20.9	21.9	10.5	9.8	11.7	13.0
Connecticut	13.5	14.1	15.6	17.4	8.0	7.3	8.5	9.3
Delaware	19.4	17.7	14.0	15.8	8.8	10.3	11.8	11.1
District of Columbia	23.6	30.4	17.3	19.5	8.6	5.8	8.4	12.1
Florida	13.7	15.4	17.2	18.1	9.5	7.6	10.6	10.1
Georgia	18.5	19.5	18.9	20.5	11.2	8.7	11.7	11.3
Hawaii	18.4	20.9	17.6	19.2	5.9	6.9	7.7	8.4
Idaho	22.3	22.7	25.5	26.6	14.4	7.4	14.5	10.3
Illinois	15.1	16.5	15.1	17.5	7.3	6.9	9.7	8.7
Indiana	19.3	19.9	18.4	20.4	11.3	10.1	11.8	12.4
Iowa	17.6	18.9	19.6	19.4	6.6	7.5	9.8	9.5
Kansas	19.4	20.7	22.0	24.7	10.8	9.9	12.8	12.6
Kentucky	17.3	18.2	19.9	22.4	11.0	10.6	15.0	13.3
Louisiana	18.0	18.4	19.4	21.9	6.8	8.3	10.4	10.4
Maine	16.4	15.0	16.4	16.9	11.0	10.7	11.9	12.5

Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia (Continued)

State	Marriage Rates				Divorce Rates			
	2014		2008		2014		2008	
	Women	Men	Women	Men	Women	Men	Women	Men
Maryland	14.7	16.0	18.2	20.4	8.2	6.8	10.0	10.1
Massachusetts	14.3	14.6	13.4	15.1	7.0	6.5	8.8	7.9
Michigan	14.1	14.9	15.7	16.7	9.2	8.5	9.9	10.3
Minnesota	16.9	17.9	18.0	18.7	7.8	7.6	9.3	9.3
Mississippi	19.2	20.7	19.1	21.6	10.0	11.3	11.6	14.0
Missouri	16.7	17.2	18.8	21.2	8.4	9.7	11.3	11.2
Montana	16.6	16.5	18.4	18.7	6.9	12.2	10.3	9.6
Nebraska	16.7	18.0	23.8	25.3	8.9	8.6	9.7	12.3
Nevada	21.2	21.3	21.9	21.6	10.4	9.8	11.2	12.8
New Hampshire	18.6	18.5	15.7	17.5	8.8	9.1	9.2	9.8
New Jersey	13.7	15.9	15.4	17.5	5.5	6.6	8.7	7.0
New Mexico	16.9	17.8	21.7	23.0	9.7	8.5	10.2	9.6
New York	14.4	16.7	14.7	17.1	6.3	5.9	7.1	7.0
North Carolina	17.1	18.8	18.5	21.1	9.0	7.8	10.1	9.5
North Dakota	18.5	18.8	23.9	26.3	7.6	5.0	6.0	10.2
Ohio	15.2	16.1	16.5	18.0	8.7	7.8	9.9	9.8
Oklahoma	22.6	23.5	23.0	24.4	12.5	11.6	16.5	13.6
Oregon	18.3	18.4	19.3	20.5	10.4	8.7	11.3	10.4
Pennsylvania	15.0	16.2	14.3	16.0	7.2	7.2	8.0	8.2
Rhode Island	13.9	16.6	13.9	16.5	8.6	6.5	8.5	8.4

Item 1: Table of Marriage and Divorce Rates per 1,000 Women and Men Aged 15 or Older in Each U.S. State and the District of Columbia (Continued)

State	Marriage Rates				Divorce Rates			
	2014		2008		2014		2008	
	Women	Men	Women	Men	Women	Men	Women	Men
South Carolina	16.8	19.5	15.4	15.7	8.9	8.9	9.8	9.5
South Dakota	21.1	22.7	17.2	21.9	7.6	7.1	8.9	11.2
Tennessee	18.6	19.8	18.2	20.3	9.5	11.0	12.8	12.8
Texas	19.5	21.1	20.6	22.2	10.2	9.8	12.7	10.9
Utah	26.4	24.3	28.4	27.6	11.9	9.5	14.0	12.1
Vermont	14.9	15.2	14.7	15.8	6.7	6.7	10.1	11.6
Virginia	18.9	19.6	18.5	21.0	8.7	8.1	11.2	9.7
Washington	19.8	19.6	21.0	22.5	8.7	8.8	12.5	11.5
West Virginia	17.0	18.1	17.4	19.6	12.4	9.5	12.8	10.3
Wisconsin	16.4	16.1	18.4	18.8	7.9	7.5	9.8	8.5
Wyoming	26.4	24.6	28.1	27.5	8.1	6.8	10.3	21.0

Note: These are survey data and do not necessarily reflect marriages and divorces that took place in a particular state. Instead, they reflect whether people living in a particular state at the time of the American Community Survey interview had divorced or married within the 12 months prior to the interview.