Morse Code, Scrabble, and the Alphabet.

Mary Richardson  
John Gabrosek  
Diann Reischman  
Phyllis Curtiss  
Grand Valley State University

Journal of Statistics Education Volume 12, Number 3 (2004),  
ww2.amstat.org/publications/jse/v12n3/richardson.html

Copyright © 2004 by Mary Richardson, John Gabrosek, Diann Reischman, and Phyllis Curtiss, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Influential observation; Outlier; Regression assumptions; Regression diagnostics; Simple linear regression.

Abstract

In this paper we describe an interactive activity that illustrates simple linear regression. Students collect data and analyze it using simple linear regression techniques taught in an introductory applied statistics course. The activity is extended to illustrate checks for regression assumptions and regression diagnostics taught in an intermediate applied statistics course.

1. Introduction

In this paper, we discuss a data collection and analysis activity for illustrating simple linear regression and outlier analysis. The activity was designed to involve students in the process of data collection and to motivate studying the relationship between two quantitative variables. Students collect data on occurrences of letters in English text. These data are used to study the relationships between how often a letter occurs in English text, and: (1) the letter’s Morse Code units and (2) the relative frequency of Scrabble™ game tiles for the letter.

We typically use this activity to illustrate and reinforce simple linear regression concepts that have been previously discussed in class lectures. In addition, we use this activity to introduce students to the use of a statistical software package for performing simple linear regression analysis.

The main activity consists of two stand-alone parts. The instructor may choose to use both parts or only one part of the main activity. Beyond the main activity, we have included two extensions. The first extension has students check the assumptions for simple linear regression. The second extension has students perform outlier analysis. We have designed the main activity and the two extensions so that they may be used as stand-alone activities or in sequence. Use of a computer software package or statistical calculator is required for Part I and Part II of the activity. Use of a computer software package is required for the extensions.
2. Data Collection

Each student is given a copy of the Data Collection Sheet (see Appendix A) and asked to find a magazine, newspaper, or World Wide Web article consisting of at least 300 letters. As homework, students use their articles to obtain empirical distributions for the relative frequency of the letters of the alphabet in English text. Students complete the table on the Data Collection Sheet. During one class period, a class distribution is generated by dividing students into small groups, having them obtain group distributions, and tallying the group results. The next class period students write the class distribution onto the tables on Worksheet 1 (see Appendix B). We have included a typical example of a class distribution in Table 1 that we use for illustration throughout this paper. The calculations done throughout this paper use the class relative frequencies from Table 1 rounded to the nearest hundredth.

Table 1. Pooled letter frequencies for one class.

| Letter | Class Total | Class Relative Frequency | Letter | Class Total | Class Relative Frequency |
|--------|-------------|--------------------------|--------|-------------|--------------------------|
| A      | 631         | 8.41                     | O      | 546         | 7.28                     |
| B      | 100         | 1.33                     | P      | 170         | 2.27                     |
| C      | 240         | 3.20                     | Q      | 8           | 0.11                     |
| D      | 285         | 3.80                     | R      | 457         | 6.09                     |
| E      | 912         | 12.16                    | S      | 526         | 7.01                     |
| F      | 148         | 1.97                     | T      | 685         | 9.13                     |
| G      | 163         | 2.17                     | U      | 216         | 2.88                     |
| H      | 344         | 4.59                     | V      | 70          | 0.93                     |
| I      | 553         | 7.37                     | W      | 151         | 2.01                     |
| J      | 19          | 0.25                     | X      | 20          | 0.27                     |
| K      | 63          | 0.84                     | Y      | 153         | 2.04                     |
| L      | 317         | 4.23                     | Z      | 10          | 0.13                     |
| M      | 179         | 2.39                     | Total  | 7500       | 100                      |

Rather than have students collect articles and tally data on letter relative frequencies, the instructor could provide an empirical distribution for letter relative frequencies (Sinkov 1980; Malkevitch and Froelich 1993). Another possibility would be to have students make use of an online resource containing written text, such as Project Gutenberg at www.promo.net/pg/, and use an online letter counter to obtain letter relative frequencies. One such letter counter can be found at ww2.amstat.org/publications/jse/secure/v7n2/count-char.cfm

3. Part I: Relative Frequency of Letter versus Morse Code Units

3.1 Background
We ask students to explore the relationship between a letter’s relative frequency in English text and the letter’s International Morse Code units. According to the web site www.arrl.org/FandES/ead/learncw/, in 1832, Samuel Morse conceived the basic idea of an electromagnetic telegraph. Experiments with various kinds of electrical instruments and codes resulted in a demonstration of a working telegraph set in 1836, and introduction of the circuit relay. This made transmission possible for any distance. With his creation of the American Morse Code, the historic message, “What hath God wrought?” was successfully sent from Washington to Baltimore in 1844.

Morse Code originated on telegraph lines and the original users did not listen to tones but instead to clicking sounds created by sounders. They used the American Morse Code as opposed to today’s International Morse Code. The Morse Code unit is a measure of the length of time required to transmit a signal. A **dit**, represented in our text by a dot (·), has a value of one Morse Code unit. A **dah**, represented in our text by a dash (-), has a value of three Morse Code units. The space between two components in a sequence of dits and dahs has a value of one Morse Code unit. For example, the Morse Code for the letter “A” is (· -). This equates to 5 Morse Code units: 1 for the dit, 1 for the space, and 3 for the dah.

Both the American Morse Code and International Morse Code use the same principle: the most common letters have the shortest codes. In order to determine the incidence of each letter Morse went to his local newspaper. There he found compositors making up pages by hand from individual letters. Morse simply counted the number of pieces of type for each letter, thinking that this must be related to the number needed. Thus “E” has the shortest code, ‘dit’, whereas “Z” is ‘dah-dah-dit-dit’ and “Q” is ‘dah-dah-dit-dah’. It is interesting to note that the symbol for “V”, ‘dit-dit-dit-dah’, is also the opening phrase of Beethoven’s Fifth (V’th) Symphony. Morse was 20 years younger than Beethoven - was he a fan of the composer? (Reference: www.rod.beavon.clara.net/morse.htm.)

After we explain the background of the origin and development of Morse Code, we ask students to think about what relationship there may be between how often letters of the English alphabet appear in printed media, and the corresponding Morse Code units for the letters.

### 3.2 Procedure

Each student needs a copy of Worksheet 1 (Appendix B). Students write the class letter frequency distribution obtained earlier onto **Table 1 of Worksheet 1**. Given a letter’s International Morse Code, students calculate the Morse Code units. This is done for all letters except “G” and “H”, whose Morse Code units will be predicted using simple linear regression.

Worksheet 1 contains a series of questions that students must answer. Students are asked to state whether they think that the association between a letter’s Morse Code units and the letter’s relative frequency in English text is positive or negative. Then, they construct a scatterplot (Figure 1) with the Morse Code units on the vertical (y) axis and the letter relative frequency on the horizontal (x) axis. Students see that there is a negative, linear association between letter relative frequency and Morse Code units. The exercise of guessing the direction of a relationship between two quantitative variables is a useful skill for students. It helps them to connect what they see in the scatterplot with the logical relationship between a letter’s relative frequency in English text and its Morse Code units.

The letter “O”, which has atypically high Morse Code units for its relative frequency, is identified as a bivariate outlier. American Morse Code was widely used in land-line communications in which the signals were carried across land by lines (wires) supported by telegraph poles. American Morse Code was well suited for land-line communication but could not easily be used for radio telegraphic communication due to embedded spaces, which were actually an integral part of several letters. In particular, the letter “O” was ‘dah-space-dah’ in the American Morse Code. The International Morse Code eliminated all of the embedded spaces and long dashes within letters that were found in many of the letters in the American Morse Code, and the letter “O” became ‘dah-dah-dah’. (Reference: chss.montclair.edu/~pererat/telegraph.html.)
Figure 1. Scatterplot of International Morse Code Units versus Relative Frequency of Letter in English Text (based on the class data shown in Table 1)

Next, students find the linear correlation $r$. The correlation, without using the letters “G” and “H”, is $r = -0.82$. Students discover that even for a value of $r$ rather close to -1, there is noticeable spread of the data about a linear pattern.

Next, students sketch a “fit-by-eye” line to the scatterplot. Often the fit is very poor. Students tend to fit the line to the extremes or to outliers rather than to the majority of the points. Different students will have different “fit-by-eye” lines for the same data. Students are led to the need for an objective criterion for finding a “best” line.

Students find the least-squares regression line as the “best” line. The correct equation for the line is $\hat{y} = 11.38 - 0.81x$. Since the data used to fit the regression line do not include “G” or “H”, students use the equation to predict the Morse Code units for these two letters. Given the true Morse Code units for “G” and “H”, students calculate the residuals. The predicted Morse Code units for “G” and “H” are 9.62 and 7.66, respectively. The actual Morse Code units for “G” and “H” are 9 (- - ·) and 7 (· · · ·), respectively.

By comparing the regression predictions for “G” and “H” to the value of $\bar{y} = 8.25$ students see that the information contained in an explanatory variable may give insight into the value of the response variable.

### 4. Part II: Relative Frequency of Letter versus Relative Frequency of Scrabble Tile

#### 4.1 Background
Scrabble is a board game in which players use letters on tiles (one letter per tile) to construct words. Each letter is assigned a certain number of points, and when a player uses that letter he/she receives the assigned number of points. Letters that are harder to use are worth more points. For example, use of the letter “Q” results in 10 points for the player, while use of the letter “E” results in only 1 point. In a standard Scrabble game set, there are 100 total letter tiles. Two of these letter tiles are blanks that can be substituted for any letter. The remaining 98 tiles are distributed to the 26 letters of the alphabet. How popular is the Scrabble board game? According to Hasbro, it can be found in an estimated one of every three American homes. The following history of Scrabble was taken from the web site www.hasbroscrabble.com.

Who invented Scrabble? During the Great Depression, an out-of-work architect named Alfred Mosher Butts decided to invent a board game. He did some market research and concluded that games fall into three categories: number games, such as dice and bingo; move games, such as chess and checkers; and word games, such as anagrams. Butts wanted to create a game that combined the vocabulary skills of crossword puzzles and anagrams, with the additional element of chance.

How did he do it? Butts studied the front page of The New York Times to calculate how often each of the 26 letters of the English language was used. He discovered that vowels appear far more often than consonants, with “E” being the most frequently used vowel. After figuring out frequency of use, Butts assigned different point values to each letter and decided how many of each letter would be included in the game. The letter “S” posed a problem. While it is frequently used, Butts decided to include only four “S” tiles in the game, hoping to limit the use of plurals. After all, he didn’t want the game to be too easy!

The boards for the first versions of the game were hand drawn with architectural drafting equipment, reproduced by blueprinting and pasted on folding checkerboards. The tiles were similarly hand-lettered, then glued to quarter-inch balsa and cut to match the squares on the board.

Butts’ first attempts to sell his game to established game manufacturers were failures. He and his partner, entrepreneur James Brunot, refined the rules and design of the game, and named it Scrabble. The name, which means “to grope frantically,” was trademarked in 1948. As so often happens in the game business, Scrabble plugged along, gaining slow but steady popularity among a comparative handful of consumers. Then in the early 1950s, as legend has it, the president of Macy’s discovered the game while on vacation, and ordered some for his store. Within a year, everyone “had to have one,” and Scrabble sets were being rationed to stores around the country.

After we explain the background of the origin and development of the board game Scrabble, we ask students to think about what relationship there may be between how often letters of the English alphabet appear in printed media, and the corresponding percentage of Scrabble tiles allotted to the letters.

4.2 Procedure

Each student needs a copy of Worksheet 1 (Appendix B). Worksheet 1 contains a series of questions that students must answer. Students write the class letter frequency distribution obtained earlier onto Table 2 of Worksheet 1. Students are asked to state whether they think that the association between a letter’s Scrabble tile relative frequency and the letter’s relative frequency in English text is positive or negative. Students then construct a scatterplot with the Scrabble tile relative frequency on the vertical (y) axis and the relative frequency in English text on the horizontal (x) axis (Figure 2). There is a positive, linear association between a letter’s relative frequency in English text and the letter’s Scrabble tile relative frequency. Students are confirmed in their expectation that the most frequent letters in English text have the most Scrabble tiles.
Next, students find the linear correlation $r$. The correlation, without using the letters “L” and “W”, is $r = 0.92$. Then, students sketch a “fit-by-eye” line to the scatterplot. Different students will have different “fit-by-eye” lines to the same data. Students are led to the need for an objective criterion for finding a “best” line.

Students use least-squares regression to find the “best” line. The correct equation for the line is $\hat{y} = 0.54 + 0.86x$. The data used to fit the regression line does not include “L” or “W”, so students can use the equation to predict the Scrabble tile relative frequency for these two letters. Given the true Scrabble tile relative frequencies for “L” and “W”, students calculate the residuals. The predicted Scrabble tile relative frequencies for “L” and “W” are 4.18 and 2.27, respectively. The actual Scrabble tile relative frequencies for “L” and “W” are 4.08 and 2.04, respectively. Students see that the regression equation is very good at predicting the frequency of Scrabble tiles.

Note to the Instructor: (Information taken from: www.askoxford.com/asktheexperts/faq/aboutwords/frequency?view=uk.) To determine the frequency of the letters of the alphabet in English, both Morse and Butts essentially counted the number of occurrences of each letter in English text. However, English text is dominated by a relatively small number of common words such as “the”, “of”, “and”, “a”, “to”, and so on. An analysis of the letters occurring in the words listed in the main entries of the Concise Oxford Dictionary (9th edition, 1995) produced the values shown in Table 2 (table entries have been rounded):

| Letter | Frequency |
|--------|-----------|
| E      | 11.2%     |
| M      | 3.0%      |

Note to the Instructor: (Information taken from: www.askoxford.com/asktheexperts/faq/aboutwords/frequency?view=uk.) To determine the frequency of the letters of the alphabet in English, both Morse and Butts essentially counted the number of occurrences of each letter in English text. However, English text is dominated by a relatively small number of common words such as “the”, “of”, “and”, “a”, “to”, and so on. An analysis of the letters occurring in the words listed in the main entries of the Concise Oxford Dictionary (9th edition, 1995) produced the values shown in Table 2 (table entries have been rounded):
The third column gives the ratio of a letter’s frequency to that of the letter “Q.” The letter “E” is over 56 times more common than “Q” in forming individual English words.

| Letter | Relative Frequency | Scrabble Tile Value |
|--------|--------------------|---------------------|
| A      | 8.5%               | 43.3                |
| R      | 7.6%               | 38.6                |
| I      | 7.5%               | 38.4                |
| O      | 7.2%               | 36.5                |
| T      | 7.0%               | 35.4                |
| N      | 6.7%               | 33.9                |
| S      | 5.7%               | 29.2                |
| L      | 5.5%               | 28.0                |
| C      | 4.5%               | 23.1                |
| U      | 3.6%               | 18.5                |
| D      | 3.4%               | 17.2                |
| P      | 3.2%               | 16.1                |
| H      | 3.0%               | 15.3                |
| G      | 2.5%               | 12.6                |
| B      | 2.1%               | 10.6                |
| F      | 1.8%               | 9.2                 |
| Y      | 1.8%               | 9.1                 |
| W      | 1.3%               | 6.6                 |
| V      | 1.0%               | 5.1                 |
| K      | 1.1%               | 5.6                 |
| J      | 0.2%               | 1.0                 |
| Q      | 0.2%               | 1.4                 |
| Z      | 0.3%               | 1.4                 |
| X      | 0.3%               | 1.5                 |
| E      | (1)                |                     |

In the context of Scrabble, for example, given that the game requires players to form words, it might seem logical to make the Scrabble tile letter ratios nearer to those occurring in different words rather than the actual frequency of each letter in English text. If, for example, “Q” occurs 1/57 as often in English words than “E” does, then maybe “Q” should be nearer to 1/57 as frequent as “E” in the Scrabble tiles. However, this is not what Butts did when developing the game. Therefore, we have students find the relative frequencies of letters in English text rather than relative frequencies of letters in a listing of English words.

**Note to the Instructor:** Students might also be asked to explore the relationship between a letter’s relative frequency in English text and the *Scrabble tile point value* for that letter. Figure 3 shows a scatterplot with the Scrabble tile point value on the vertical axis and the letter’s relative frequency on the horizontal axis. There is a curved association between a letter’s relative frequency and its Scrabble tile points.
5. Extension 1: Checking Regression Assumptions

In this extension, we ask students to check the required assumptions for simple linear regression (listed on Worksheet 2 in Appendix C). We emphasize to students that these assumptions make it possible for us to develop measures of reliability for the least-squares estimators and to develop hypothesis tests for examining the usefulness of the least-squares line. We also note that there are various techniques for checking the validity of the assumptions and that there are remedies to apply when the assumptions appear not to hold. We emphasize that the regression assumptions need not hold for the least-squares model to give accurate predictions.

Our goal in using this extension is to introduce a general discussion of residual analysis and to have students use statistical software to perform a residual analysis.

Each student needs a copy of Worksheet 2 (Appendix C). This worksheet is a continuation of Worksheet 1, Part II. In Section 4, we described an exploration of the relationship between a letter’s relative frequency in English text and the letter’s Scrabble tile relative frequency. Students were asked to construct a scatterplot, describe the association, and find the least-squares regression line. In this extension, we include a basic residual analysis to check the simple linear regression assumptions. We base our residual analysis on the data shown in Table 1. The letters “L” and “W” are again omitted, as on Worksheet 1, Part II.

We begin by giving students a brief introduction to residual analysis. Using the least-squares regression line found in Part II of Worksheet 1, students compute the predicted values and residuals for the Scrabble tile relative frequencies of the 24 letters that were used to obtain the regression line. Students plot the residuals (vertical axis) against the predicted values (horizontal axis) and interpret the plot. In particular, we ask students to discuss what the plot may indicate about the appropriateness of the simple linear regression model. Is there an apparent pattern on the plot? Or, does the plot show an unstructured horizontal band of
points centered at zero? Figure 4 shows the residual plot.

An examination of the residual plot reveals an increasing trend or a “cone” shape of the residual variability, which implies that the constant error variance assumption may be violated. We discuss with students that one way to stabilize the variance of the random errors may be to refit the model using a transformation on the independent or dependent variable or both variables. We then introduce and discuss a square root transformation, with the transformed model given by: \( \sqrt{y} = \beta_0 + \beta_1 + \varepsilon \). We ask students to transform the value of the dependent variable (Scrabble tile relative frequency), fit the transformed model to the data, and construct a new residual plot. Figure 5 shows the residual plot for the transformed model. For a detailed discussion on data transformations see Mendenhall and Sincich (1996) or Neter, Kutner, Nachtsheim, and Wasserman (1996).
When interpreting the plot of the transformed-model residuals, students will notice that the square root transformation resulted in dampening the increasing residual trend and that the new residual plot more closely resembles the ideal of an unstructured horizontal band of points centered at zero.

Students check for nonnormal errors by constructing a Q-Q plot of the transformed residuals and examining the plot to see if a linear pattern is displayed. **Figure 6** shows the Q-Q plot of the transformed residuals.
The points on the Q-Q plot do not reveal a substantial deviation from a linear pattern, indicating that there is no reason to assume that the error distribution is not a normal distribution.

Because this may be the first exposure many students have had to transformations, we ask students to use the transformed least-squares regression line to predict the relative frequency of Scrabble tiles for the letters “L” and “W”. The equation of the transformed least-squares regression line is given by $\sqrt{y} = 1.03 + 0.21x$. We explain to students that we are now predicting the square root of the relative frequency of Scrabble tiles, and that in order to predict the relative frequency of Scrabble tiles, we must square the predicted value that we obtain. The data used to fit the transformed regression line does not include “L” or “W”, so students can predict the Scrabble tile relative frequency for these two letters. Given the true Scrabble tile relative frequencies for “L” and “W”, students calculate the residuals. The predicted Scrabble tile relative frequencies for “L” and “W” are 3.68 and 2.11, respectively. The actual Scrabble tile relative frequencies for “L” and “W” are 4.08 and 2.04, respectively. Students see that the transformed regression equation is very good at predicting the frequency of Scrabble tiles.

In the second extension, which follows in Section 6, we give an in-depth discussion of regression diagnostics and the determination of influential points.

6. Extension 2: Outlier Analysis

Each student needs a copy of Worksheet 3 (Appendix D). This worksheet is a continuation of Worksheet 1, Part I. In Section 3, we described an exploration of the relationship between a letter’s relative frequency in English text and the letter’s Morse Code units. Students constructed a scatterplot, described the association, and found the least-squares regression line. In this extension, we include a discussion of regression diagnostics and the role of influential points. We base our regression diagnostic calculations on the data.
shown in Table 1. The letters “G” and “H” are again omitted, as on Worksheet 1, Part I. Many of the calculations on Worksheet 3 require the use of a computer software package such as SPSS. We refer the reader to Appendix D to check for any notation that is not familiar and to Appendix E for detailed solutions to the Worksheet 3 questions.

We use the two-dimensional case as a bridge to multiple linear regression where the use of regression diagnostics and the analysis of influential points is important to a full understanding of the model. Students benefit from first encountering these ideas in the familiar case of simple linear regression.

We begin by having students show that the regression line passes through the point \( \left( \overline{x}, \overline{y} \right) = (3.85, 8.25) \) for the regression of \( y = \) Morse Code units on \( x = \) letter relative frequency and that, in general, the regression line passes through \( \left( \overline{x}, \overline{y} \right) \). Students see that the regression line tilts about the point \( \left( \overline{x}, \overline{y} \right) \) and that a slope of 0 corresponds to the line \( y = \overline{y} \) where the value of the \( x \) variable provides no useful information about the value of the \( y \) variable.

Next, students are asked to find the sum of squares for error (\( SSE = 84.30 \)) and the root mean square error \( s = \sqrt{\frac{84.30}{24 - 2}} = 1.96 \). Students choose a point to remove from the data set that would decrease the root mean square error. The root mean square error is calculated after that point has been removed, and its value is compared to the original value. Students should choose a letter that falls far from the regression line, for instance the bivariate outlier “O”. Students see that outliers increase \( s \).

The coefficient of determination, \( r^2 \), is examined. Students are asked to find and interpret the value of the coefficient of determination for the application. Next, students are asked to remove a letter that would increase the value of \( r^2 \). The value of \( r^2 \) is calculated with the point removed and its value is compared to the overall \( r^2 \). Again letter “O” would be a logical choice.

The rest of the questions focus on diagnostic measurements in regression analysis. The letter “O” is an outlier in the data set, and the letter “E” follows the general pattern of the data although it has an atypically high relative frequency of occurrence in English text. These two letters are used as examples to compare diagnostic values for these two types of observations. For a detailed discussion of regression diagnostics, see Neter, Kutner, Nachtsheim, and Wasserman (1996).

Students are asked to find and interpret the studentized residuals for the letters “O” and “E” (\( r_O = 2.93 \) and \( r_E = -0.35 \)). Letter “O” has a large positive studentized residual suggesting that it may be a positive outlier. This is consistent with the scatterplot that students constructed earlier.

The studentized residual for a point includes all points that were used in the calculation of the regression equation. If the point is an outlier then the root mean square error will be biased upwards and the studentized residual for that point will be pulled toward 0. The externally studentized residual (R-student) is calculated for a given point without including the point in the calculation of the regression equation. Students are asked to find and interpret R-student for letters “O” and “E” (\( t_O = 3.67 \) and \( t_E = -0.34 \)).

We use DFFITS, DFBETAS, and Cook’s D to ascertain a point’s actual influence on predicted values and the regression equation. Students are asked to find and interpret these statistics for the letters “O” and “E”.

The standardized DFFITS for letters “O” and “E” are 1.12 and –0.23, respectively. Thus, the estimated number of standard errors for the fitted value for letter “O” would increase by more than 1 standard error if “O” were excluded from the regression. Students see that letter “O” has a large influence on its fitted value.
The standardized DFBETAS measure the influence that a point has on the $y$-intercept and the slope, separately. For letter “O”, the standardized DFBETAS are $DFBETAS_{y\text{-int}} = -0.11$ and $DFBETAS_{slope} = 0.81$. For letter “E”, the standardized DFBETAS are $DFBETAS_{y\text{-int}} = 0.11$ and $DFBETAS_{slope} = -0.21$. Thus, letter “O” has a large influence on the slope but not the $y$-intercept, while letter “E” does not have undue influence on either of the regression coefficients. Students are now able to see exactly where the effect of the outlying point is felt in the regression equation.

Cook’s D is a measure of the overall influence that a point has on the regression equation. For letters “O” and “E”, Cook’s D is 0.40 and 0.03, respectively. The Cook’s D value for the letter “O” is large when compared to the other Cook’s D values in the data set (next largest is .085 for the letter “Y”). Thus, letter “O” is having a very large influence on the regression equation.

After students have completed the questions in this extension, we give a summary discussion of the regression diagnostics calculated for the letters “O” and “E”. In addition, we discuss whether or not the letter “O” should be removed from the regression equation. We note that if the goal is to use the regression equation for prediction (say of the Morse Code units for letters “G” and “H”), then it might be better to exclude letter “O” when finding the regression equation.

7. Conclusions

This activity has a wide range of possible uses and extensions. It can be used in lower level undergraduate courses as well as in intermediate level undergraduate courses. We use the activity to provide introductory statistics students with an opportunity to practice the application of simple linear regression analysis in two applied problems. In an intermediate course, we use the activity to illustrate checking simple linear regression assumptions and as an introduction to regression diagnostics and the role of influential points.

The activity is received very well by students. Most students are familiar with the Scrabble board game and have some understanding of Morse Code. We begin the activity with a discussion of the development of Morse Code and a reminder of how the Scrabble game is played. In addition, we discuss how Alfred Butts came up with the letter tile distribution for Scrabble. These discussions create an interest on the part of the students. Students are further interested when we explain that we are going to collect data in order to come up with our own letter usage distribution. And, they enjoy examining the relationship between the class distribution and the distributions developed by Morse and Butts. Constructing scatterplots allows students to visualize how strongly the class distribution correlates to the distributions developed by Morse and Butts. Calculating least-squares regression lines and using them to make predictions shows students the usefulness of least-squares regression for predicting the value of a response variable.

Another very strong point of the activity is that it does not require the use of a substantial amount of in-class time. In addition, no extra materials are required to use this activity.

Appendix A: Data Collection Sheet

Instructions:

Obtain a newspaper, magazine, or www article. Within the article, choose a place to begin at random and tally the letters one at a time, filling out the Individual Tally column in the table provided below. Add up the tallies, which should come to a grand total of at least 300 letters (or more, if desired). Then, for each letter, enter the individual frequency. Save the newspaper or magazine article that you use. Underline the portion of the article from which you obtained your tallies. Turn in your article along with your Data Collection...
The purpose of this activity is to examine the relationship between a letter’s relative frequency in English text and the letter’s Morse Code units.

**Instructions:** The Morse Code unit is a measure of the length of time required to transmit a signal. The duration of a dit (·) is one unit and the duration of a dah (-) is three units. The space between the components
of the sequence of dits and dahs for a letter is one unit. For example, the Morse Code for the letter “A” is (· -). This equates to 5 Morse Code units: 1 for the dit, 1 for the space between the dit and the dah, and three for the dah. The table below gives the class distribution of the letters of the alphabet in English text and the Morse Code for each letter except for “G” and “H”. Begin by finding the Morse Code units for each letter.

| Letter | Estimated Relative Frequency in English Text | International Morse Code | Morse Code Units | Letter | Estimated Relative Frequency in English Text | International Morse Code | Morse Code Units |
|--------|---------------------------------------------|---------------------------|-----------------|--------|---------------------------------------------|---------------------------|-----------------|
| A      | 8.41                                        | · -                       |                 | N      | 7.12                                        | - ·                       |                 |
| B      | 1.32                                        | - · · ·                   |                 | O      | 7.28                                        | - - -                     |                 |
| C      | 3.20                                        | - · ·                      |                 | P      | 2.27                                        | · - -                     |                 |
| D      | 3.80                                        | - ·                        |                 | Q      | 0.11                                        | -- · -                    |                 |
| E      | 12.16                                       | ·                          |                 | R      | 6.09                                        | · - ·                     |                 |
| F      | 1.97                                        | · · -                      |                 | S      | 7.01                                        | · · ·                     |                 |
| G      | 2.17                                        | ?                          |                 | T      | 9.13                                        | -                        |                 |
| H      | 4.59                                        | ?                          |                 | U      | 2.88                                        | · · -                     |                 |
| I      | 7.37                                        | · ·                        |                 | V      | 0.93                                        | · · · -                   |                 |
| J      | 0.25                                        | · · · ·                    |                 | W      | 2.01                                        | · - -                     |                 |
| K      | 0.84                                        | - · -                      |                 | X      | 0.27                                        | · · · -                   |                 |
| L      | 4.23                                        | · - ·                      |                 | Y      | 2.04                                        | - · -                     |                 |
| M      | 2.39                                        | - -                        |                 | Z      | 0.13                                        | - · ·                     |                 |

**Questions:**

1. Do you think that the association between a letter’s Morse Code units and the letter’s relative frequency in English text will be positive or negative? Why?

2. Make a scatterplot with each letter’s Morse Code units on the vertical (y) axis and the letter’s relative
frequency in English text on the horizontal (x) axis.

3. Are there any letters whose Morse Code units do not follow the pattern for the majority of points? If so, which letters are they? (Such points are called bivariate outliers.)

4. Find the linear correlation $r$ between a letter’s relative frequency in English text and the letter’s Morse Code units.

5. Using the scatterplot and $r$, describe the pattern, strength, and direction of the association between a letter’s relative frequency in English text and the letter’s Morse Code units.

6. Remove one letter from the scatterplot that would result in increasing the value of $r$ for the 23 remaining letters. Which letter did you remove? Why? What is the new value of $r$?

7. Sketch a “fit-by-eye” line on your scatterplot. Find the equation of your “fit-by-eye” line.

8. Why is a “fit-by-eye” approach not the best way to find the line?

9. Find the equation of the least-squares regression line.

10. Plot your least-squares line on your scatterplot.

11. Compare and contrast your “fit-by-eye” line to the least-squares regression line.

12. Interpret the slope of the least-squares regression line.

13. Use your least-squares regression line to predict the Morse Code units for the letters “G” and “H”. Complete the following table.

| Letter | Estimated Relative Frequency in English Text | Morse Code Units | Predicted Morse Code Units | Residual |
|--------|---------------------------------------------|------------------|----------------------------|----------|
| G      | 2.17                                        | 9 (- - ·)        |                            |          |
| H      | 4.59                                        | 7 (· · · ·)       |                            |          |

14. Notice that the predicted Morse Code units for “G” and “H” from the regression equation are closer to the actual Morse Code units for “G” and “H” than is the mean Morse Code units for all the letters excluding “G” and “H”. Given this, does knowing a letter’s relative frequency in English text provide useful information as to its Morse Code units?

**Part II:**

The purpose of this activity is to examine the relationship between a letter’s relative frequency in English text and the relative frequency of Scrabble tiles for the letter.

**Instructions:**

The table below gives the class distribution of the letters of the alphabet in English text and the relative frequency of Scrabble tiles containing the letter (except for “L” and “W”). Using this information, we wish to examine the relationship between a letter’s relative frequency in English text and the relative frequency of Scrabble tiles for the letter.
| Letter | Estimated Relative Frequency in English Text | Relative Frequency of Scrabble Tiles | Letter | Estimated Relative Frequency in English Text | Relative Frequency of Scrabble Tiles |
|--------|------------------------------------------|-----------------------------------|--------|------------------------------------------|-----------------------------------|
| A      | 8.41                                     | 9.18                              | N      | 7.12                                     | 6.12                              |
| B      | 1.32                                     | 2.04                              | O      | 7.28                                     | 8.16                              |
| C      | 3.20                                     | 2.04                              | P      | 2.27                                     | 2.04                              |
| D      | 3.80                                     | 4.08                              | Q      | 0.11                                     | 1.02                              |
| E      | 12.16                                    | 12.24                             | R      | 6.09                                     | 6.12                              |
| F      | 1.97                                     | 2.04                              | S      | 7.01                                     | 4.08                              |
| G      | 2.17                                     | 3.06                              | T      | 9.13                                     | 6.12                              |
| H      | 4.59                                     | 2.04                              | U      | 2.88                                     | 4.08                              |
| I      | 7.37                                     | 9.18                              | V      | 0.93                                     | 2.04                              |
| J      | 0.25                                     | 1.02                              | W      | 2.01                                     | ?                                 |
| K      | 0.84                                     | 1.02                              | X      | 0.27                                     | 1.02                              |
| L      | 4.23                                     | ?                                 | Y      | 2.04                                     | 2.04                              |
| M      | 2.39                                     | 2.04                              | Z      | 0.13                                     | 1.02                              |

Questions:

1. Do you think that the association between relative frequency of Scrabble tiles and relative frequency of the letter in English text will be positive or negative? Why?

2. Make a scatterplot with each letter’s relative frequency of Scrabble tiles on the vertical (y) axis and the letter’s relative frequency in English text on the horizontal (x) axis.

3. Are there any letters whose relative frequency of Scrabble tiles does not follow the pattern for the majority of points? If so, which letters are they?

4. Find the linear correlation \( r \) between a letter’s relative frequency in English text and the relative
frequency of Scrabble tiles for the letter.

5. Using the scatterplot and \( r \), describe the pattern, strength, and direction of the association between a letter’s relative frequency in English text and the relative frequency of Scrabble tiles for the letter.

6. Sketch a “fit-by-eye” line on your scatterplot.

7. Find the equation of your “fit-by-eye” line.

8. Find the equation of the least-squares regression line.

9. Plot your least-squares line on your scatterplot.

10. Compare and contrast your “fit-by-eye” line to the least-squares regression line.

11. Interpret the slope of the least-squares regression line.

12. Use your least-squares regression line to predict the relative frequency of Scrabble tiles for the letters “L” and “W”. Complete the following table.

| Letter | Estimated Relative Frequency in English Text | Relative Frequency of Scrabble Tiles | Predicted Relative Frequency of Scrabble Tiles | Residual |
|--------|---------------------------------------------|-------------------------------------|-----------------------------------------------|----------|
| L      | 4.23                                        | 4.08                                |                                               |          |
| W      | 2.01                                        | 2.04                                |                                               |          |

---

Appendix C: Worksheet 2

Checking assumptions for the regression of \( y = \) Scrabble Tile Relative Frequency on \( x = \) Relative Frequency of Letter in English Text.

For this Worksheet, the use of a computer software package such as SPSS or Minitab is required to carry out the calculations and obtain the graphs. All of the calculations on this Worksheet are done without the letters “L” and “W”.

Recall that the straight-line probabilistic model is given by: \( y = \beta_0 + \beta_1 x + \varepsilon \).

In a simple linear regression analysis, we must make four basic assumptions about the general form of the probability distribution of the random error \( \varepsilon \).

1. The mean of the probability distribution of \( \varepsilon \) is 0.

2. The variance of the probability distribution of \( \varepsilon \) is constant for all settings of the independent variable \( x \).

3. The probability distribution of \( \varepsilon \) is normal.

4. The values of \( \varepsilon \) associated with any two observed values of \( y \) are independent. That is, the value of \( \varepsilon \)
associated with one value of \( y \) has no relation with the values of \( \mathcal{E} \) associated with other \( y \) values.

Because the assumptions concern the random error component, \( \mathcal{E} \), of the model, the first step is to estimate the random error associated with each \( x \) value. This estimated error is called the regression residual and is denoted by \( \hat{\mathcal{E}} \). A residual, \( \hat{\mathcal{E}} \), is defined as the difference between an observed \( y \) value and its corresponding predicted value: 

\[
\hat{\mathcal{E}} = (y - \hat{y}) = y - \left( \hat{\beta}_0 + \hat{\beta}_1 x \right).
\]

The residual can be calculated and used to estimate the random error and to check the regression assumptions. Such checks are generally referred to as residual analyses. The residuals should look like they are a random sample from a normal population with a mean of 0 and a constant variance.

**Goal:** We want to determine if a simple linear regression model is an appropriate model for describing the relationship between Scrabble Tile Relative Frequency and Relative Frequency of a Letter in English Text.

1. In general, we can check for a misspecified model and/or non-constant error variance by plotting the residuals against the predicted values, \( \hat{y} \). The plot should show an unstructured horizontal band of points centered at zero. If the plot shows a pattern, this may indicate that the model is misspecified or that the error variance is not constant.

   Using the least-squares regression line found in Part II of Worksheet 1, compute the predicted values and residuals for the Scrabble tile relative frequencies of the 24 letters that were used to obtain the regression line (that is, all letters except for “L” and “W”). Plot the residuals (\( y \)-axis) against the predicted values (\( x \)-axis). Interpret the plot and discuss what the plot may indicate about the appropriateness of the simple linear regression model. (Is there an apparent pattern in the plot, or does the plot show an unstructured horizontal band of points centered at zero?)

2. If an examination of the residuals plotted against the predicted values reveals a “cone” shape of the residual variability (the size of the residuals increases as the predicted Scrabble tile relative frequencies increase), this implies that the constant error variance assumption is violated. One way to stabilize the variance of the random errors may be to refit the model using a transformation on the independent or dependent variable or both variables. When the residual plot shows a cone-shaped pattern, one useful variance-stabilizing transformation is the square root transformation, with the transformed model given by: 

\[
\sqrt{y} = \beta_0 + \beta_1 x + \varepsilon.
\]

   a. Transform the value of the dependent variable (Scrabble tile relative frequency) and construct a scatterplot of the transformed data. Interpret the scatterplot.

   b. Fit the transformed regression model (\( \sqrt{y} = \beta_0 + \beta_1 x + \varepsilon \)) to the data.

3. For the transformed regression model, perform a residual analysis.

   a. Plot the residuals (\( y \)-axis) against the predicted values (\( x \)-axis). Interpret the plot and discuss what the plot indicates about the appropriateness of the transformed model.

   b. Are there any letters that have atypically large residuals?

   c. Construct a Q-Q plot of the residuals. Does the Q-Q plot indicate that the residuals do not have a normal distribution? Explain your answer.

4. Use the transformed least-squares regression line to predict the relative frequency of Scrabble tiles for the letters “L” and “W”. Complete the following table.
Appendix D: Worksheet 3

Regression Diagnostics for the Regression of $y = \text{Morse Code Units}$ versus $x = \text{Relative Frequency of Letter in English Text}$.

The use of a computer software package such as SPSS or Minitab is necessary to carry out many of the calculations in this Worksheet. All of the calculations in this Worksheet are done without the letters “G” and “H”.

1. The equation of the least-squares regression line is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_0$ is the $y$-intercept and $\hat{\beta}_1$ is the slope.
   
   a. Find the equation of the least-squares line.

   b. Verify that the least-squares line passes through the point $(\bar{x}, \bar{y})$.

   c. The $y$-intercept can be found as $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Using this fact, show that the least-squares line passes through the point $(\bar{x}, \bar{y})$ for any regression. (Hint: Show that when $x = \bar{x}$ then $\hat{y}$ must equal $\bar{y}$.)

   d. If the slope of the least-squares regression line is 0, then what is the equation of the line? (Hint: Substitute correct values for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$).

   e. What does a line with slope 0 tell you about the usefulness of the value of the $x$ variable for predicting the value of the $y$ variable?

2. The estimated error variance $s^2$ is a measure of the variability between the actual $y$-coordinates ($y$) and the predicted $y$-coordinates $\hat{y}$. It is obtained by dividing the sum of squares of residuals

   $$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \text{ by (n-2).}$$

   (We use $n-2$ because we must estimate the $y$-intercept and the slope of the least-squares line in order to find $\hat{y}$. We lose one degree of freedom for each quantity estimated.)

   The value $s^2$ is often called the Mean Square Error (MSE).

   You may be able to obtain an intuitive feeling for $s$ by remembering that the least-squares line estimates the mean value of $y$ for a given value of $x$. Because $s$ measures the spread of the distribution of the $y$ values about the least-squares line, we expect that at least 75% (according to Chebyshev’s Theorem) of the observed $y$ values will lie within $2s$ of their respective least squares predicted values.

| Letter | Estimated Relative Frequency in English Text | Relative Frequency of Scrabble Tiles | Predicted Relative Frequency of Scrabble Tiles | Residual |
|--------|---------------------------------------------|-------------------------------------|-----------------------------------------------|----------|
| L      | 4.23                                        | 4.08                                |                                                |          |
| W      | 2.01                                        | 2.04                                |                                                |          |
1. Calculate the SSE.

2. Calculate $s$. Interpret its value using Chebyshev’s Theorem and two standard deviations.

3. Remove one letter from the scatterplot that would result in decreasing the value of $s$ for the 23 remaining letters. What letter did you remove? Why? What is your new value for $s$?

3. The coefficient of determination $r^2$ is the proportion of the sum of squares of deviations of the $y$ values about $\bar{Y}$ (i.e., a measure of the variability in the $y$ values) that can be attributed to a linear relationship between $y$ and $x$. The value of $r^2$ is always between 0 and 1.

a. Find and interpret $r^2$.

b. Remove one letter from the scatterplot that would result in increasing the value of $r^2$ for the 23 remaining letters. What letter did you remove? Why? What is your new value for $r^2$?

4. Ideally each observation will have the same influence on the regression equation. Unfortunately, this is never the case in practice. And that’s okay, provided that no observations exert an undue influence on the regression equation. There are many measures of an observation’s influence on a regression equation. One such measure is the studentized residual $r_i$ of the $i^{th}$ observation. The studentized residuals are unit-free and $t$-like. A large $|r_i|$ (say $\geq 2.5$) suggests that the observation may be an outlier.

a. Find and interpret $r_i$ for letter “O”.

b. Find and interpret $r_i$ for letter “E”.

5. A potential problem with using the studentized residual to detect whether the $i^{th}$ observation is an outlier is that the calculation of $r_i$ includes the $i^{th}$ observation. If the $i^{th}$ observation is an outlier then the value of $r_i$ will be biased toward 0 and so it will become more difficult to detect outlier points. The externally studentized residual (often called R-student) $t_i$ of the $i^{th}$ observation is calculated without including the $i^{th}$ observation. The externally studentized residuals are also $t$-like in behavior and so $|t_i| \geq 2.5$ suggests the observation may be an outlier.

a. Find and interpret $t_i$ for letter “O”.

b. Find and interpret $t_i$ for letter “E”.

6. Both $r_i$ and $t_i$ are useful for determining the influence of an observation on errors in the $y$-direction. They help us to detect outliers. We are also concerned with the influence of an observation on the regression equation. We do not want just a few observations to be almost solely responsible for the $y$-intercept or slope of the line. One measure of the influence of an observation on a regression equation is the statistic standardized DFFITS. It is used to determine the influence that the $i^{th}$ observation has on the regression equation.
on the fitted value \( \hat{y}_i \). Standardized DFFITS is a measure of the estimated number of standard errors that the fitted value for the \( i^{th} \) point changes if the point is excluded from the regression. Most statisticians frown upon using a specified cutoff for what values of DFFITS suggest that an observation is unduly influential. Instead they compare the DFFITS for all the observations and look for any values that appear to be much greater than the majority of the points.

a. Find and interpret standardized DFFITS for the letter “O”.

b. Find and interpret standardized DFFITS for the letter “E”.

7. The statistic standardized DFBETAS measures the number of standard errors that a particular regression coefficient changes if the observation is not included. A large value in magnitude of standardized DFBETAS \( j,i \) indicates that the \( i^{th} \) point exerts strong influence on the \( j^{th} \) coefficient. For a simple linear regression the \( y \)-intercept is coefficient \( j = 0 \) and the slope is coefficient \( j = 1 \). Some statisticians consider \( \text{DFBETAS} > \frac{2}{\sqrt{p}} \) to be large. Usually, a point will have a large standardized DFBETAS only if it has a large standardized DFFITS. The standardized DFFITS identifies the observation as influential and the standardized DFBETAS identifies which regression coefficients are heavily influenced by the observation.

a. Find and interpret standardized DFBETAS \( 0,i \) for the letter “O”.

b. Find and interpret standardized DFBETAS \( 0,i \) for the letter “E”.

c. Find and interpret standardized DFBETAS \( 1,i \) for the letter “O”.

d. Find and interpret standardized DFBETAS \( 1,i \) for the letter “E”.

8. While DFBETAS measures an observation’s influence on each regression coefficient individually, Cook’s Distance (or Cook’s D) statistic measures the combined influence of an observation on all regression coefficients. As with DFFITS, a large Cook’s D is evaluated against the other observations in the data set.

a. Find and interpret \( D_i \) for the letter “O”.

b. Find and interpret \( D_i \) for the letter “E”.

Appendix E: Answers to Worksheet 3 Questions

1. a. The equation of the least-squares line is \( \hat{y} = 11.38 - 0.81x \).

b. When \( x = 3.88 \), \( \hat{y} = 11.38 - 0.81(3.88) = 8.24 \).

c. \( \hat{y} = \hat{\beta}_0 + \hat{\beta}_1x \Rightarrow \hat{y} = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1x \). When \( x = \bar{x} \Rightarrow \hat{y} = \bar{y} \).
d. \[ \hat{y}_1 = 0 \Rightarrow \hat{y}_0 = \bar{y} \Rightarrow \hat{y} = \bar{y}. \]

e. If \( \hat{\beta}_1 = 0 \), then the \( x \) variable provides no information that is useful for predicting the value of \( y \).

2.

a. \( \text{SSE} = 84.30 \).

b. \( s = \sqrt{\frac{84.30}{22}} = 1.96 \). We expect at least 75% of the Morse Code units to be within 3.92 of their predicted value \( \hat{y} \).

c. Answers will vary depending on what letter was selected, but the letter “O” will be a common choice. Removing letter “O” would decrease \( s \) because “O” has a large squared residual. The value for \( s \) for the regression without “O” is \( s = \sqrt{\frac{51.34}{21}} = 1.56 \).

3.

a. \( r^2 = 0.699 \). This means that 66.9% of the variation in Morse Code units is explained by the linear relationship with letter frequency in English text.

b. Answers will vary depending on what letter was selected, but the letter “O” will be a common choice. Removing letter “O” would increase \( r^2 \) because it has a large squared residual. The value of \( r^2 \) for the regression without “O” is \( r^2 = 0.792 \).

4.

a. The studentized residual for letter “O” is \( r_O = 2.93 \). Because \( r_O \) is a large positive value, we conclude that letter “O” may be an atypically large positive outlier. Letter “O” does not follow the pattern of the rest of the data.

b. The studentized residual for letter “E” is \( r_E = -0.35 \). Because \( r_E \) is close to 0, we conclude that letter “E” is not an outlier. Letter “E” follows the pattern of the rest of the data.

5.

a. The externally studentized residual for letter “O” is \( t_O = 3.67 \). Because \( t_O \) is quite far from 0 we conclude that “O” is probably an outlier and does not follow the pattern of the rest of the data.

b. The externally studentized residual for letter “E” is \( t_E = -0.34 \). Because \( t_E \) is close to 0 we conclude that “E” is not an outlier and does follow the pattern of the rest of the data.

6.

a. For letter “O”, standardized DFFITS is 1.12. This means that if we exclude “O” from the regression then the fitted value for “O” will change by more than 1 standard error. The next largest standardized DFFITS value is 0.43, thus the value for “O” would be considered unusually large in this data set.

b. For letter “E”, standardized DFFITS is –0.23. This means that if we exclude “E” from the regression then the fitted value for “E” will change by less than 1/4 of a standard error. Eleven
of the 24 letters have a larger standardized DFFITS in absolute value, thus this value is not unusual in this data set.

7. a. For letter “O”, standardized DFBETAS<sub>0,O</sub> on the y-intercept is –0.11. Because
\[ \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{24}} = 0.408 \]
this would imply that “O” has very little influence on the y-intercept.

b. For letter “E”, standardized DFBETAS<sub>0,E</sub> on the y-intercept is 0.11. Because
\[ \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{24}} = 0.408 \]
this would imply that “E” has very little influence on the y-intercept.

c. For letter “O”, standardized DFBETAS<sub>1,O</sub> on the slope is 0.81. Because
\[ \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{24}} = 0.408 \]
this would imply that “O” has a large influence on the slope.

d. For letter “E”, standardized DFBETAS<sub>1,E</sub> on the slope is –0.21. Because
\[ \frac{2}{\sqrt{n}} = \frac{2}{\sqrt{24}} = 0.408 \]
this would imply that “E” has only marginal influence on the slope.

8. a. For letter “O”, Cook’s D is 0.40. The next largest Cook’s D value is .085, thus the Cook’s D value for “O” would be considered to be a large value so we conclude that “O” has a significant influence on the regression equation overall.

b. For letter “E”, Cook’s D is 0.03. Eleven of the 24 letters have a larger Cook’s D value, thus we conclude that “E” has little influence on the regression equation overall.

Acknowledgments

The authors gratefully acknowledge the helpful comments and suggestions of the editor and the referees during the preparation of this manuscript.

References

Malkevitch, J. and Froelich, G. (1993), “Loads of Codes,” in HistoMAP, Module 22.

Mendenhall, W., and Sincich, T. (1996), A Second Course In Statistics: Regression Analysis (5th ed.), New Jersey: Prentice Hall.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. (1996), Applied Linear Statistical Models (4th ed.), IRWIN: Chicago.
Sinkov, A. (1980), *Elementary Cryptanalysis: A Mathematical Approach in the New Mathematical Library, Number 22*, Mathematical Association of America.

Mary Richardson  
Department of Statistics  
Grand Valley State University  
1 Campus Drive  
Allendale, MI 49401  
U.S.A.  
*richamar@gvsu.edu*

John Gabrosek  
Department of Statistics  
Grand Valley State University  
1 Campus Drive  
Allendale, MI 49401  
U.S.A.  
gabrosej@gvsu.edu

Diann Reischman  
Department of Statistics  
Grand Valley State University  
1 Campus Drive  
Allendale, MI 49401  
U.S.A.  
reischmd@gvsu.edu

Phyllis Curtiss  
Department of Statistics  
Grand Valley State University  
1 Campus Drive  
Allendale, MI 49401  
U.S.A.  
curtissp@gvsu.edu