Coefficient of Determination:

It’s Just Middle School Math?

Susan A. Peters
University of Louisville
s.peters@louisville.edu

Rose Mary Zbiek
The Pennsylvania State University
rmz101@psu.edu
**Coefficient of Determination: It's Just Middle School Math?**

Statistics content occupies a prominent position in curriculum expectations. Recommendations for statistics teaching include actively engaging students with important statistical ideas and appropriately integrating technology into activities (Garfield & Ben-Zvi, 2008). Teachers can benefit from statistical learning opportunities that align with these recommendations (Chance & Rossman, 2006). In this article, we share experiences from our design and implementation of an activity to develop conceptual understanding of the coefficient of determination, $R^2$. We use it to provide prospective teachers with a learning opportunity aligned with statistics teaching recommendations and encourage its use with secondary students. The activity offers a unique combination of opportunities to use technology for doing and learning statistics, to draw on calculus ideas, and to make surprising connections to secondary school mathematics content.

The coefficient of determination activity helps teachers and students develop multiple aspects of bivariate data analysis—analysis of data involving two variables. This area of statistics is prominent in curriculum recommendations, including *Focus in High School Mathematics* (NCTM, 2009), *Principles and Standards for School Mathematics* (NCTM, 2000), and the *Common Core State Standards in Mathematics* (CCSSM [CCSSI, 2010]). *Common Core State Standards in Mathematics* additionally emphasizes modeling and strategic use of appropriate tools. Recommendations for mathematics teacher preparation (AMTE, 2006; CBMS, 2012; Franklin et al., 2015) parallel these curriculum themes. We designed the coefficient of determination activity to target these recommendations and to encourage flexible thinking about statistics content.
Coefficient of Determination Activity

The coefficient of determination activity is about moving beyond knowing the coefficient of determination as a number for $R^2$ to reasoning about it as a concept in the context of least squares regression to model relationships between variables. The activity was designed to develop conceptual understanding of $R^2$, model relationships between variables and assess goodness of fit, make connections to common mathematics content (and among representations), and promote technology as a tool for learning and teaching. Technology that allows for dynamic interaction with data, such as the freely available Core Math Tools from the National Council of Teachers of Mathematics, is best for this activity.

Prior to Data-Based Exploration

Within their activity experience, teachers first encounter $R^2$ by reading Gloria Barrett’s (2000) Mathematics Teacher article about how her students explored the meaning of the concept. Her definition of $R^2$ is “the proportion of variation in the response variable that is accounted for by a regression model and any explanatory variables with which it is associated” (p. 230). She also notes several properties of $R^2$, including its relationship to correlation and the importance of values of 0 and 1.

After reading the article, teachers work in groups to write a description of $R^2$ in a way that could be understood by secondary students not familiar with the concept. Examples of groups’ descriptions appear in Figure 1.
Figure 1. Teachers’ initial descriptions of $R^2$

We do not discuss the correctness or completeness of descriptions, but instead discuss commonalities among them. Underlined passages in Figure 1 identify coefficient of determination, relationship, ideas about 0 and 1, and independent and dependent variables as elements common across the class’ descriptions. By examining common passages, teachers note important aspects of the concept to consider while working through the data-based exploration in the appendix. They revisit these descriptions after the exploration.

Also, prior to the exploration, we ask teachers to solve the Software Purchase problem shown in Figure 2. The class returns to this seemingly unrelated problem and its purpose after the exploration. We encourage the reader to solve the Software Purchase problem and to work through the exploration before reading further.
You’ve become so excited by the prospect of converting music file formats with Switch that you search the Internet to find the lowest purchase price for this software. You notice that you can order the student version of Switch from NCH Software (http://www.nch.com.au/switch/index.html) for approximately $40. When you check eBay (http://www.ebay.com), you find a seller who is offering the same package for approximately $32. What percent of NCH Software’s list price do you save when you buy from the eBay seller?

Figure 2. Software Purchase problem

**Data-Based Exploration**

Teachers continue exploring $R^2$ by engaging in an exploration based on the data and general idea of Barrett’s (2000) activity. The exploration tasks motivate bivariate analyses and draw attention to connections among a broad range of representations. We also update the type of technology used in the activity to capitalize on dynamic environments.

Data for the exploration are the ages of husbands and wives for 16 couples shown in Table 1. Teachers have access to dynamic statistics software and to a file containing the data. They work in pairs to respond to the tasks shown on the exploration sheet.
Table 1. Ages of husbands and wives for 16 couples (Barrett 2000)

<table>
<thead>
<tr>
<th>Wives’ Ages</th>
<th>Husbands’ Ages</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>45</td>
</tr>
<tr>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>45</td>
<td>43</td>
</tr>
<tr>
<td>44</td>
<td>50</td>
</tr>
<tr>
<td>52</td>
<td>42</td>
</tr>
<tr>
<td>41</td>
<td>52</td>
</tr>
<tr>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>40</td>
<td>39</td>
</tr>
<tr>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>40</td>
<td>51</td>
</tr>
<tr>
<td>34</td>
<td>37</td>
</tr>
<tr>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>37</td>
<td>43</td>
</tr>
<tr>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>45</td>
<td>48</td>
</tr>
<tr>
<td>38</td>
<td>41</td>
</tr>
</tbody>
</table>

**Univariate Setting**

The first task ask teachers to predict the age of a 17th husband using only the 16 husband’s ages, dotplots (Figure 3), or some other representation. Most teachers draw on their prior experiences with elementary statistics and predict an age using a measure of center. They estimate the median age from the dotplot, calculate the mean age, or use the modes of 42 to 44 and the data’s slight rightward skew to estimate an age in the interval 43 to 45.
Teachers intuitively if not explicitly suggest that measures of center or small intervals of values around measures of center provide the best possible predictions when data for a single variable is the only information available. We ask them how confident they are in predicting the age of a 17th husband using only husband’s ages. Their responses suggest low levels of confidence as they turn to the real-world context to provide real or hypothetical examples that challenge their predictions. We ask them what might increase confidence, including what additional information might be productive.

**Bivariate Setting**

Teachers’ explorations with the bivariate data begin with their constructing a scatterplot of the response variable of husband’s ages versus the explanatory variable of wives’ ages and a horizontal line drawn through the mean age of the husbands (Figure 4). They consider how well the average husband age predicts the age of a 51-year-old wife’s husband. We purposefully chose a wife’s age that is not represented in the data and for which the average husband age does not appear to provide a good prediction. The value predicted from the line does not follow the flow of data displayed in the scatterplot. We ask teachers to compare the value predicted by the mean line with their predictions from the univariate setting. They see that the use of the mean in the bivariate setting considers only husbands’ ages. Teachers quickly recognize the fact that, in this bivariate setting, they can make a better prediction for the husband’s age by considering wives’ ages. The question we then turn to becomes *how best* to use this additional information.
Fig 4. Scatterplot of husband’s ages versus wives’ ages and horizontal line that represents the average husband’s age

**Lines of “Best” Fit**

To incorporate the wives’ ages, most teachers suggest using linear regression to fit a line to the data and to predict the age of the 51-year-old wife’s husband. They use a graph (Figure 5) and the upward trend of data to justify why the regression line yields a better fit and prediction than the mean line. More difficult to justify is why the regression line is better than, for example, a line with a slightly greater slope and slightly lesser y-intercept. Dynamic features of technology offer one opportunity to explore why.

Figure 5. Scatterplot of husband’s ages versus wives’ ages and least squares line
Dynamic technology allows teachers to add a moveable line to the scatterplot and to adjust the line’s slope and y-intercept to fit numerous lines to the data “by hand” (Figure 6). By examining many lines with slightly different coefficients, teachers notice the need for common criteria so that everyone can talk about and make predictions using the same line. At this point, they have a need for a common method to fit lines to data but not a justification of why least squares regression tends to be the chosen method. Examining the variety of criteria used to fit lines tends to elicit a common criterion that the line should be as close as possible to as many data points as possible. To facilitate further discussion, we introduce residual and define residual in this context as the difference between an observed husband’s age and the husband’s age predicted from a line fit to the data.

![Graph](image)

Figure. 6. Moveable line with an adjustable intercept (♀) and slope (♂) with squares shown

Multiple criteria for “best fit” emerge, and the discussion focuses on how and why those criteria produce a best fitting line. In general, the criteria involve minimizing residual values in some way. One suggestion might be to minimize the sum of residual values. This general process, however, does not yield a unique line. Any line that passes through the point with an
abscissa of the mean of the wives’ ages and an ordinate of the mean of the husband’s ages yields a residual sum of zero [e.g., \( \text{Husband \_ Age} = 0.25(\text{Wife \_ Age} - 41.25) + 44.6875 \) and \( \text{Husband \_ Age} = -0.25(\text{Wife \_ Age} - 41.25) + 44.6875 \)]. To eliminate the effects of signed residual values, teachers typically suggest minimizing the sum of the absolute values of the residuals. Some teachers recall an analogous situation with standard deviation and suggest minimizing the sum of squared residuals, which also results from a sum of positive values. We use both suggestions to make a connection to calculus.

**Connections to Calculus**

Teachers recognize optimization problems from calculus. They expect to define a function and find maximum or minimum values using derivatives. The question they undertake is: Which function? If we were to use squared residuals, we need to minimize a function associated with a quadratic expression: \( (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \ldots + (y_n - \hat{y}_n)^2 \), where \( y_i \) represents the age of husband \( i \) and \( \hat{y}_i \) represents the predicted age of the husband. Finding the derivative in a quadratic situation seems feasible. If we use absolute deviations, we would need to find a minimum for a function arising from an expression with absolute value: \( |y_1 - \hat{y}_1| + |y_2 - \hat{y}_2| + \ldots + |y_n - \hat{y}_n| \). The discontinuous nature of this function’s derivative makes it difficult to find a minimum value. The function also does not produce a unique minimum and thus, does not yield a unique line—a violation that led us into our discussion about “best fit.”

To explain minimizing the sum of squared residuals, we again turn to dynamic technology and show squares on the scatterplot with the moveable line (Figure 6). Each square is formed from a data value and using the residual for that data value as the length of a side of the square (shown for the mean line and the least squares line in Figure 7a and Figure 7b, respectively). Teachers connect the area of each square to the squared residual value for the
corresponding data value and the sum of the areas to the sum of squared residuals. As teachers move the line, the displays for the equation of the line and the sum of squared residuals automatically update. The sum of squares from each moveable line is less than the sum of squares from the least squares line—empirical verification that the least squares line produces the minimum sum of squared residuals.

Fig. 7. Area representation for the mean line (a) and for the least squares line (b)

**Linking Representations**

Connections between the graphs shown in Figure 7 require explicit attention to elements of the graphs. It helps to superimpose the green square on the blue square related to one data point. In Figure 8a, the side lengths for the green and blue square correspond to the residual from the least squares line and the mean line, respectively. Comparing the areas of the two squares shows that, for this point, the square of the least squares line residual value is less than the square of the mean line residual value. Teachers make similar comparisons for several data points (Figure 8). They conclude that, for these data, the area increases for the squares associated with some of the points but decreases for a majority of data values to produce the least sum in the case
of the least squares line. Drawing connections between graphical representations precedes linking the graphical representations to the symbolic formula for calculating $R^2$.

$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$.

We challenge teachers to explain how the symbols in the formula connect to the graphs. Teachers observe that $y_i - \bar{y}$ is the residual value using the horizontal line for the mean age of the husbands. The square of this value is the area of a blue square with side lengths corresponding to the residual value, and $\sum(y_i - \bar{y})^2$ yields the sum of the areas. Similarly, $y_i - \hat{y}_i$ yields corresponding values from the least squares line for a green square. The difference between the two expressions yields the sum of the “extra” areas. Dividing by $\sum(y_i - \bar{y})^2$ produces the ratio of the extra area to the total area for the squares based on the mean line.

**Connections to Middle School Mathematics**

To further solidify understandings of $R^2$, the class returns to the Software Purchase problem (Figure 2) and their solutions, such as those in Figure 9. Teachers observe that most if not all solutions involve subtracting 32 from 40, dividing by 40, and converting a fraction to a
percent. We ask teachers to compare their approaches with the approach shown in Figure 10 and we identify the Software Purchase problem as a “percent decrease” problem like those most students see in middle school mathematics.

Figure. 9. Fraction/percent (a), equation (b), and proportion (c) solutions to Software Purchase problem

Figure. 10. Targeted solution to Software Purchase problem

We next consider how the percent decrease problem connects to the coefficient of determination. Some people react immediately, observing, “that’s $R^2$-squared!” Others need to analyze further the formula for calculating $R^2$ and the targeted percent decrease solution. The parallel nature of the ratios can be seen by examining the expressions:
When teachers compare the two expressions, they observe that the mean line area sum corresponds to the initial price and the sum of squared residuals corresponds to the price paid. From this link they conclude that the ratio of the difference of the least squares area and the mean line area is the percent of the mean line area that is accounted for by the least squares area. These comparisons lead teachers to interpret $R^2$ as the percent decrease in variation in the response variable (husband_age) from the regression model.

Similar connection of formula and area representations arises as teachers consider a second formula for $R^2$, $R^2 = \frac{SST - SSE}{SST}$. Within this equation, SST represents the total sum of squares or the sum of squared residuals for the mean line, and SSE is the sum of squared errors, which is the sum of the areas of the squares for the graph (Figure 7b).

**Return to Definition of Coefficient of Determination**

To bring closure to the activity, we ask teachers to revise their initial descriptions of the coefficient of determination. We find that teachers use many of the same words, but they now move beyond paraphrasing Barrett’s written words. For example, they are able to describe what is meant by “proportion of variability” as a percent by using the area representations or the symbolic formula for calculating $R^2$.

Additionally, familiarity with the graphical and symbolic representations allows teachers to determine the conditions under which $R^2$ assumes a value of 0 or 1. A value of 0 results when $\sum(y_i - \hat{y}_i)^2$ is equal to $\sum(y_i - \bar{y})^2$ and the least squares line is the mean line. Rewriting

$$\frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{40 - 32}{40}$$
\[
\frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \text{ as } 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2},
\]
teachers see that a value of 1 results if the value of \(\sum (y_i - \hat{y}_i)^2\) is equal to 0; the least squares line provides a perfect fit to the data.

**Closing Comments**

We designed the coefficient of determination activity to help teachers develop conceptual understanding of the coefficient of determination by making connections between graphical and symbolic representations, to illustrate the use of technology to support statistics learning, and to make connections across statistics and mathematics content. Critical to their enhanced understanding is seeing \(R^2\) as the percent of variability in the response variable (husband_age in the exploration) accounted for by the regression model. Teachers connect statistics and mathematics content as they explicitly link standard symbolic forms and graphical representations. They connect \(R^2\) to the middle grades notion of percent decrease and use their knowledge of differentiation methods to reason about why the least squares line is a unique “best” fitting line. In addition, the activity exemplifies the use of dynamic technology for learning and doing statistics.
References


http://www.amte.net/publications.


Appendix

Data-Based Exploration Inspired by Barrett (2000)

Open the dynamic statistics file named Barrett_data.

Examining a Univariate Data Distribution

Create a dotplot of the Husband_Age variable by creating a new graph and dragging the Husband_Age variable to the horizontal axis. Use the information to answer the following questions.

1. The ages for sixteen husbands are displayed in the dotplot. If the age of a seventeenth husband was collected but not recorded, what age would you guess for the missing age and why?

2. How confident are you that the age you guessed matches the actual age of the husband? To what do you attribute your level of confidence?

Examining Bivariate Data Distributions and the Coefficient of Determination, $R^2$

Create a scatterplot of the Wife_Age and Husband_Age variables using Wife_Age on the horizontal axis and Husband_Age on the vertical axis. Plot the function Husband_Age = mean(Husband_Age) on the scatterplot.

3. Describe how well the average husband age predicts the age of a 51 year-old wife’s husband.

Choose the “Show squares” option from the Graph menu. Create a second scatterplot of the Wife_Age and Husband_Age variables. Add the least-squares regression line using the Graph menu and again choose the “Show squares” option.

4. Using the least squares regression line, describe how well the line predicts the age of a 51 year-old wife’s husband.

5. Which prediction, the prediction from 3 or from 4, allows you to predict the husband’s age with more confidence and why?

6. How do the squares in the original graph relate to the formula, $R^2 = \left( \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y})^2 \right) / \sum (y_i - \bar{y})^2$?

7. How do the squares in the second graph relate to the formula for calculating $R^2$?

8. Using what you wrote for 6 and 7, describe $R^2$ in a way that someone unfamiliar with statistics could understand.