Journal of Statistics Education

Using the Height and Shoe Size Data to Introduce Correlation and Regression

Constance H. McLaren
Indiana State University

Journal of Statistics Education Volume 20, Number 3 (2012),
www.amstat.org/publications/jse/v20n3/mclaren.pdf

Key Words: Indicator Variables; Histograms; Inference.

Abstract

The Height and Shoe Size dataset contains information on height (in inches), dress shoe size, and gender for 408 college students. The information was collected to provide an interesting initial example for the study of correlation and regression in a business statistics class. Students don't mind providing this information (unlike weight, about which they are often more sensitive) and seem to enjoy seeing how accurate the resulting prediction is for their particular height or shoe size. Once each semester's values are added to the file, it is posted and used for a series of assignments. We begin with correlation, move on to a series of simple linear regression assignments, and finish by incorporating a dummy variable for gender. This data set could also be used for frequency distributions, histograms, or inference.

1. Introduction

Although text books are filled with data sets for use with correlation and regression, we have found that students are always more engaged if the data they use are closely connected to their personal experience. The students in this class are usually sophomores who have yet to enroll in business courses in finance or marketing, so data that is closer to their frame of reference makes more sense to them. We frequently use examples that deal with consumer products commonly used by college students, but we find there is even more relevance when they have supplied or collected the data themselves. One of the simplest ways to accomplish this is to send a sheet around the classroom asking each student to record height, shoe size, and gender. Entries have been saved over the years and we now have records on 408 students.

Incorporating real data, particularly that generated by the students themselves, is noted as a valuable tool for engaging students in their learning. The GAISE College report (2010) has as its second recommendation the use of real data, stating "It is important to use real data in teaching statistics to be authentic, to consider issues related to how and why the data were produced or collected, and to relate the analysis to the problem context. Using real data sets of interest to students is also a good way to engage them in thinking about the data and relevant statistical concepts." Many other authors echo this recommendation. See, for example, Auster and Wylie (2006), who wrote about approaches to active learning, Page and Mukherjee (2000) who emphasized the importance of making material relevant to the student experience, and Kottemann and Salimian (2008), whose experiment was prompted by the notion that "students rarely, if ever, have meaningful prior business experience. With the approach outlined above, students are dealing with data about themselves, which provides a context they can appreciate and that they understand, which in turn enables them to better grasp the concepts and methods being taught."

At our university, all business majors are required to complete a two-course introductory (non-calculus based) business statistics sequence, typically in their sophomore year. The first course covers data presentation, random variables and probability distributions, and inference. The second course covers tests of independence, ANOVA, correlation, regression, forecasting, and decision analysis as well as providing a brief unit on business applications of calculus. Our students are encouraged to use statistical software (e.g. JMP, Minitab, or StatCrunch), Excel, or applets to ease their calculations, but only after they have come to understand the concepts behind the calculations.

Typical business statistics texts include coverage of correlation and regression analysis (see, for example, Anderson, Sweeney, & Williams 2010; Bowerman & O'Connell 2011; Groebner, Shannon, Fry, & Smith 2011; and Levine, Berenson, Krehbiel, & Stephan 2011). Although these texts have made it a point to include large data sets for examination, we find that we are always looking for more examples. Using real data increases student interest in the topics we teach in business statistics courses, and we anticipate that this would be the case in other statistics courses as well.

Specific analytical questions for use with this data set are provided in the assignments that appear in the Pedagogical Uses section of this article. Appendix B provides solutions to these questions.

2. Data Sources

The data in the Height and Shoe Size dataset were collected over several years from college students enrolled in our second required business statistics class.

The data file is available for download in an Excel file at:
http://www.amstat.org/publications/jse/v20n3/mclaren/shoesize.xls

A documentation file describing the data set and its uses is available for download in Word format at: http://www.amstat.org/publications/jse/v20n3/mclaren/documentation.doc.

3. Description of the Data

The data file contains an index value labeled Index that counts the students and then three variables: a variable for gender (Gender, coded as F for female and M for male), dress shoe size (Size, to half sizes), and height in inches (Height, shown to two decimal places).

In the interest of time in a large class section, we chose to have students self-report their shoe sizes rather than the length of their feet. However, the use of shoe size raises issues about data level of measurement and the error introduced by sizing systems. Bowing to time constraints, we have chosen to treat the shoe size variable as numerical for the purposes of this exercise, and we consider an increase of one shoe size as, for example, going from 7 to 8 rather than from 7 to 7 1/2.

> *Potential Pitfall: A point for discussion centers on levels of data measurement. Most business statistics texts have an introductory section on nominal, ordinal, interval, and ratio data and their appropriate uses. Instructors should have students explain why shoe sizes are not ratio level data and ask them whether they can think of other examples of "doing math" with this level of data.*
>
> *Alternative Applications: If time permits, instructors might choose to have students convert their shoe size to international standards by visiting a web site such as http://www.zappos.com/measure-your-shoe-size. Alternatively, instructors who prefer that both variables be of the ratio data level could instead have each student measure and record the actual size of a bare foot. Using that measurement instead of shoe size provides ratio level data for both variables.*

Clear instructions for reporting height should be given. In the data that accompany this article, students were asked to record their height in inches; no further instructions were given. Only one student reported to the half inch, and another reported his height in centimeters (subsequently converted to inches by the instructor). It may be more practical to specify that students report and round heights to the nearest half inch.

> *Potential Pitfall: Measurement error is likely with this data. Shoe size may vary by manufacturer, and students may not have been measured recently or recall their height to more than the nearest inch. Care should be taken so that students don't assume more precision in their results than is appropriate.*

4. Pedagogical Uses

This dataset is used for a series of exercises that support the introduction to correlation and regression chapters found in many business statistics textbooks. (See, for example, the texts listed in Section 1.) We begin, on the day that the information is collected, by reviewing cross-sectional data with the students. Up until this point, most of the course work has involved a

single variable and it is important to remind students why analysts often want several measures on the same item, whether it is a store, a person, a real estate property, or something else.

After the data are collected, a file that appends the data from this class to all of the historical data (the dataset used for this article) is posted for their access. We usually sort the data by gender and then by either height or shoe size, but if it is important for your students to practice manipulating the data, you could certainly randomize the order.

> *Helpful Hint: These observations were collected from college students. If you are using this exercise for younger students or have other reasons for separating your data, you might consider including a new indicator variable to distinguish your observations from the larger set.*

Our course's study of regression begins with an introduction to correlation. The texts we use typically illustrate, via scatter plots, the meaning of positive, negative, strong, and weak linear correlation. Students are provided with the formula for the Pearson Product Moment Correlation but also use software for its calculation. They are shown a test for the significance of the correlation and also rely on the p-values that accompany software calculations of correlation.

> *Potential Pitfall: Business statistics books typically show the test statistic* $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$ *but provide little or no explanation of the assumptions required for this test. Instructors whose curriculum is more advanced will want to be sure that students understand, and check, the assumptions for the test of significance for the correlation coefficient.*

The first exercise, given as an in-class quiz or a take-home exercise, asks the students to consider correlation as preparation for learning about simple linear regression. *Note that we are a laptop campus and students have access to statistical software during class and off campus*. They are asked to calculate correlation coefficients for height and shoe size for the men and, separately, for the women, and to conduct a hypothesis test for the significance of the correlation coefficient.

Exercise 1: Correlation

Open the Height and Shoe Size Data Excel file. See Section 2 for downloading instructions. **[Local instructors may use a different name for this file.]** This file contains information from 408 college students. **[Update this total if you add observations from your students.]**

1. Create a scatter plot showing men's shoe sizes on the horizontal axis and the associated height on the vertical axis. Examine the result. Do the points suggest that the relationship between men's shoe size and height is appropriate for linear analysis? Explain your choice using several sentences.
2. What is the value of the correlation coefficient for height and shoe size for the men?
3. What is the value of the correlation coefficient for height and shoe size for the women?

4. Using either the p-value method or the critical value method, conduct this hypothesis test for the significance of the correlation of the ***women's*** heights and shoe sizes at $\alpha = 0.05$. Enter your results in **one** of the two boxes.

   $H_0: \rho = 0$
   $H_a: \rho \neq 0$

   The value of the test statistic is _____. There are _____ degrees of freedom.

Using the Critical Value Method	Using the p-value Method
The critical value of the t statistic is What is the conclusion and why?	The p-value is What is the conclusion and why?

Helpful Hint: We have found that it is worthwhile to continue to ask students to express their conclusions from hypothesis tests as full explanations and not accept a simple "reject" or "reject H_0" as the conclusion. Encourage your students to describe the test that is used, its assumptions, and the implications that result from the conclusion.

The second assignment uses the shoe size data to examine simple linear regression concepts. By the time this exercise is assigned, students have worked their way through the introductory regression analysis material in their text and understand how to interpret their results. The designation of predictor and response variables is left to the students, creating results that very clearly illustrate what parts of the regression results stay the same regardless of which variable is the predictor (x) and which is the response (y).

In this exercise we begin by considering all of the data. Students calculate correlation and compare the correlation between height and shoe size for all observations with the correlations found for each gender. The reason for aggregating the data is to motivate the use of gender as an indicator variable in the regression model.

Exercise 2: Simple Linear Regression

Open the Height and Shoe Size Data Excel file. See Section 2 for downloading instructions. **[Local instructors may use a different name for this file.]** This file contains information from 408 college students. **[Update this total if you add observations from your students.]**

1. Do you think it is reasonable to use shoe size to predict height, or is it more logical to use height to predict shoe size? Explain your choice using several sentences.
2. Using all observations, create a scatter plot showing Size on the horizontal axis and the Height on the vertical axis. If your software permits, plot the points from each gender using a different color or marker. Examine the result. Do the points suggest that the

relationship between Size and Height is appropriate for linear analysis when both genders are included? Explain your choice using several sentences.

3. Calculate the correlation coefficient for the height and shoe size variables and test its significance. Is the relationship between the two variables sufficiently strong to pursue a regression analysis? Compare the correlation coefficient for all observations to the correlation coefficient you found for a single gender in Exercise 1. To what can you attribute the fact that the correlation for all observations is larger?
4. Using the choices for the predictor (x) and response (y) variables that you indicated in question 1 develop a regression model using all the observations. Write your regression equation here using the variable names Height and Size. Explain the meaning of the values of the intercept and regression coefficient.
5. One of the students in the dataset is 64 inches tall and wears a size 8 shoe. What is the error associated with your model's prediction for this student?
6. From your analysis, determine whether a statistically significant relationship exists between Height and Size. Provide an explanation that supports your decision.
7. What percentage of the variation in the response (y) is explained by the regression model?

When the students' papers are returned, we spend some time examining the regression results from the two models they could have developed for this assignment. We find that this is a good way for students to check their understanding of the workings of regression. For those who have relied only on software and have not understood the algebraic relationships that create the ANOVA table and various other key portions of their output, we find that comparing the outputs is a revelation and assists their understanding of regression.

> *Helpful Hint: It is helpful to first have students who made different variable choices pair up to examine what is the same and what is different about their results. After they have spent some time with this discovery, call the class together and be prepared to provide both sets of results on paper or on a screen. At this point, direct students to various sum of squares formulas to see where the specification of "x" and "y" makes a difference.*

The third exercise is used after the students have learned about indicator variables. If we have a student who has worked in shoe sales, we ask for information about the way men's and women's shoes are sized. If students haven't already brought this up, this discussion supports the need for an indicator variable for gender.

Exercise 3: Indicator Variables

Open the Height and Shoe Size Data Excel file. See Section 2 for downloading instructions. **[Local instructors may use a different name for this file.]** This file contains information from 408 college students. **[Update this total if you add observations from your students.]**

1. No distinction was drawn between male and female students in Exercise 2. Rerun your model for male students alone and then for female students alone, using the same choices for dependent and independent variables that you did in Exercise 2. Are these two new models "better" than the model than combines males and females? Explain.

2. Create an indicator variable for the Gender column. You will need to decide which gender will be represented by a "1" and which by a "0." Create a multiple regression model incorporating the indicator variable and using all the observations and examine the results. Is the gender variable statistically significant? What is the expected effect of gender? Is more variation explained in this new model?
3. Use your original single variable model, your single variable gender-specific model, and your indicator variable model to predict your own measurement. Calculate the error associated with each model. Which one provided the best prediction for you? Which of these models do you believe would be best to apply to this kind of prediction in practice, the pair of gender-specific models, or the single model with gender?

   Potential Pitfall: Students may think they need to create two indicator variables, one for males and one for females. You may need to explain why only a single Gender indicator variable is needed. Remind them to keep track of which gender is coded as "1" and which as "0."

The set of three exercises provides consistency in the introduction of correlation and regression concepts. By using the same data for each of the three assignments, students begin to understand the connections between answers that often are the result of a software command.

Alternative Assignment: The exercises could be assigned individually as class coverage dictates or to a group for a project. If time is short, each exercise can stand alone. There is sufficient data that the exercises could be assigned without the addition of values from current students, but we suggest that having your students contribute their own measurements will increase interest. They may even suggest additional variables—for example, some of our students suggested adding an indicator variable for students who are NCAA athletes, believing they have relatively larger shoe sizes.

Helpful Hint: If time permits, students could be asked to develop an inventory plan for a store that sells dress shoes for both men and women and other shoes (e.g. athletic sandals, hiking boots) that could fit either men or women. Students might also be tasked with interviewing a shoe store manager or online shoe provider to understand not only the distribution of sizes in an order but also the assumed consequences for not having the correct size in stock.

5. Conclusion

The Shoe Size data set provides an interesting example to illustrate correlation, simple linear regression, and the use of indicator variables. The data set is large enough that it also generates useful residual plots. We have found that students are intrigued to see how well the models predict their results. They seem to appreciate working with the large number of observations and being included in the collection. Although we have used this data set only during coverage of correlation and regression, it could be used earlier in an introductory statistics class to illustrate frequency distributions and histograms and for confidence intervals and hypothesis testing.

Appendix A
Key to Variables in Shoe Size Data Set

For the file shoesizedata.dat (saved as tab delimited text)

Variable	Description	Label
1	Index (from 1 to 408)	Index
2	Gender (M or F)	Gender
3	Shoe size (in half sizes)	Size
4	Height (in inches)	Height

Appendix B

The Shoe Size Instructor's Manual, containing all exercise assignments and solutions, is available at http://www.amstat.org/publications/jse/v20n3/mclaren/manual.doc.

References

Anderson, D., Sweeney, D., and Williams, T. (2010), *Statistics for Business and Economics, 11th edition*, Mason, OH: Thomson South-Western.

Auster, E. R. and Wylie, K. K. (2006), "Creating Active Learning in the Classroom: A Systematic Approach," *Journal of Management Education*, 30(2), 333-353.

Bowerman, B. and O'Connell, R. (2011), *Business Statistics in Practice, 6th edition*, New York: McGraw Hill/Irwin.

Groebner, D. Shannon, P., Fry, P., and Smith, K. (2011), *Business Statistics, 8th edition*, Upper Saddle River, NJ: Pearson Education.

GAISE (Guidelines for Assessment and Instruction in Statistics Education) College Report, (2010), Zieffler, A. and Karl, S. (eds.). Available at http://www.amstat.org/education/gaise/.

Kottemann, J. E. and Salimian, F. (2008), "Engaging students in statistics." *Decision Sciences Journal of Innovative Education*, 6(2), 247–250.

Levine, D., Berenson, M., Krehbiel, T., and Stephan, D. (2011), *Statistics for Managers, 6th edition*, Upper Saddle River, NJ: Pearson Education.

Page, D. and Mukherjee, A. (2000), "Improving undergraduate student involvement in management science and business writing courses using the seven principles in action," *Education*, 120, 547-557.

Constance H. McLaren
Marketing and Operations Department
Scott College of Business
Indiana State University
Terre Haute, IN 47809
cmclaren@indstate.edu
Voice: 812.237.2282
Fax: 812.237.8129
