

## SPSS DEMONSTRATIONS [GSS18SSDS-A]

### Demonstration 1: Producing Scatterplots (Scatter Diagrams)

Despite the introduction of streaming media services (e.g., Netflix, Amazon Prime Video) and portable devices (e.g., phones and iPads), Americans are still watching television. Annually, the GSS asks respondents to report the number of hours they watch television per day. For the 2018 sample, the average number of television viewing hours was 2.97. For this SPSS Demonstration, we'll explore the relationship between *television viewing hours* (TVHOURS) and *number of years of education* (EDUC) using the techniques described in this chapter for interval-ratio data.

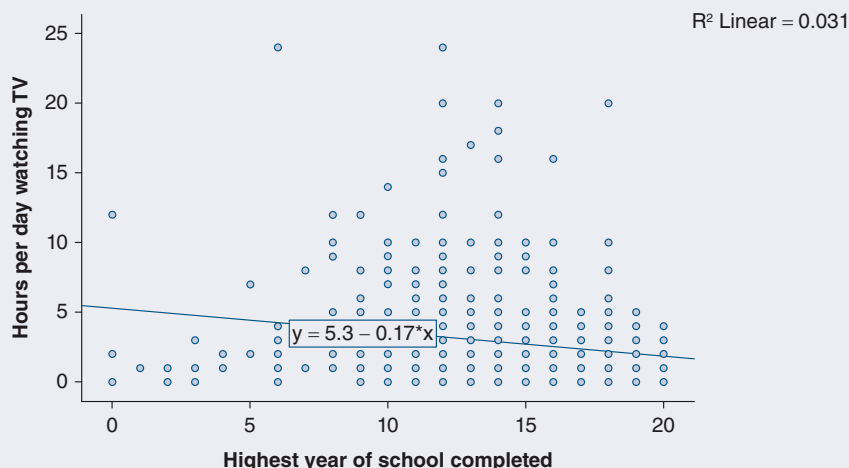
We begin by looking at a scatterplot of these two variables. The Scatter procedure can be found under the *Graphs* menu choice. In the opening dialog box, click *Legacy Dialogs*, then *Scatter/Dot* (which means we want to produce a standard scatterplot with two variables), and select the icon for *Simple Scatter*; then click *Define*.

The Scatterplot dialog box requires that we specify a variable for both the X- and Y-axes. We place EDUC (number of years of education) on the X-axis because we consider it the independent variable and TVHOURS (number of television viewing hours per day) on the Y-axis because it is the dependent variable. Then, click *OK*.

You can edit it to change its appearance by double-clicking on the chart in the viewer. The action of double-clicking displays the chart in a chart window. You can edit the chart from the menus, from the toolbar, or by double-clicking on the object you want to edit.

It is difficult to tell whether a relationship exists just by looking at points in the scatterplot, so we will ask SPSS to include the regression line. To add a regression line to the plot, we start by double-clicking on the scatterplot to open the Chart Editor. Click *Elements* from the main menu, then *Fit Line at Total*. In the section of the dialog box headed "Properties," select *Linear*. Click *Apply* and then *Close*. Finally, in the Chart Editor, click *File* and then *Close*. The result of these actions is shown in Figure 10.12.

Figure 10.12



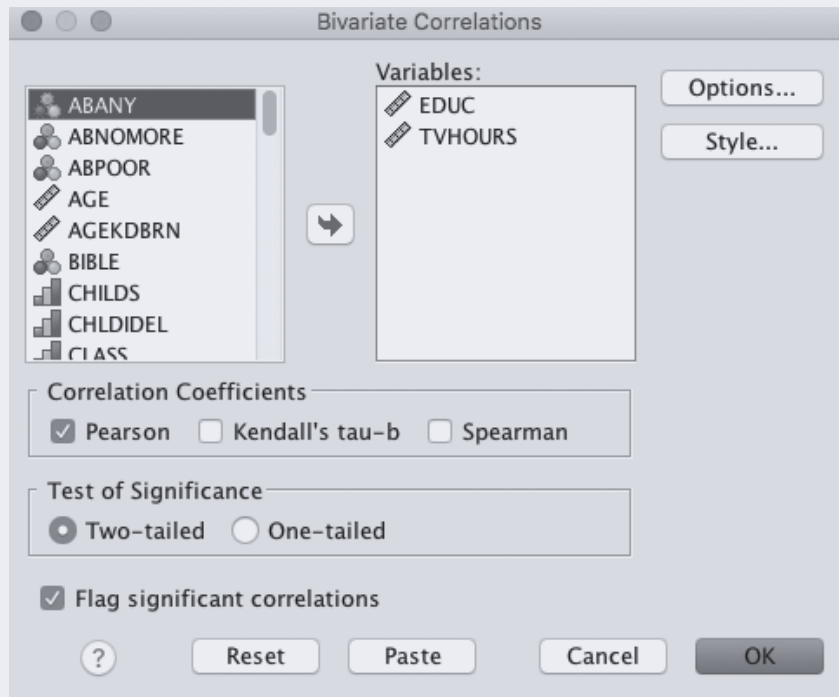
Since the regression line declines as number of years of education increases, we observe the negative relationship between education and number of television viewing hours per day. The predicted value for those with 20 years of education is about 1.9 hours, compared with 3.6 hours for those with 10 years of education. However, because there is a lot of scatter around the line (the points are not close to the regression line), the predictive power of the model is weak.

## Demonstration 2: Producing Correlation Coefficients

To further quantify the effect of education on television viewing hours, we request a correlation coefficient. This statistic is available in the Bivariate procedure, which is located by clicking on *Analyze*, *Correlate*, then *Bivariate* (Figure 10.13). Place the variables you are interested in correlating, EDUC and TVHOURS, in the Variable(s) box, then click *OK*.

SPSS produces a matrix of correlations, shown in Figure 10.14. We are interested in the correlation in the bottom left-hand cell,  $-.177$ . We see that this correlation is closer to 0 than to 1, which tells us that education is not a very good predictor of television viewing hours per day, even if it is true that those with less education work watch more hours of television per day. The number under the correlation coefficient, 1013, is the number of valid cases

Figure 10.13



**Figure 10.14**

| Correlations                     |                     |                                  |                           |
|----------------------------------|---------------------|----------------------------------|---------------------------|
|                                  |                     | Highest year of school completed | Hours per day watching tv |
| Highest year of school completed | Pearson Correlation | 1                                | -.177**                   |
|                                  | Sig. (2-tailed)     |                                  | .000                      |
|                                  | N                   | 1498                             | 1013                      |
| Hours per day watching tv        | Pearson Correlation | -.177**                          | 1                         |
|                                  | Sig. (2-tailed)     | .000                             |                           |
|                                  | N                   | 1013                             | 1014                      |

\*\* . Correlation is significant at the 0.01 level (2-tailed).

(N)—those respondents who gave a valid response to both questions. SPSS identifies significant correlations (at the .05 level or higher). The correlation between educational attainment and hours per day watching television is significant at the .01 level (two-tailed).

### Demonstration 3: Producing a Regression Equation

Next, we will use SPSS to calculate the best-fitting regression line and the coefficient of determination. This procedure is located by clicking on *Analyze, Regression*, then *Linear*. The Linear Regression dialog box (Figure 10.15) provides boxes in which to enter the dependent variable, TVHOURS, and the independent variable, EDUC (regression allows more than one). After you place the variables in their appropriate places, click *OK* to generate the output. The Linear Regression dialog box offers many other choices, but the default output from the procedure contains all that we need.

SPSS produces a great deal of output, which is typical for many of the more advanced statistical procedures in the program. The output is presented in Figure 10.16. Under the Model Summary, the coefficient of determination is labeled “R square.” Its value is .031, which is very weak. Educational attainment explains 3.1% of the variation in hours of television viewing.

The regression equation coefficients are presented in the Coefficients table. The regression equation coefficients are listed in the column headed “B.” The coefficient for EDUC, or  $b$ , is about  $-.172$ ; the intercept term, or  $a$ , identified in the “(Constant)” row, is  $5.304$ . Thus, we would predict that every additional year of education decreases the television viewing hours per day by about 10 minutes ( $-.172 \times 60$ ). Or we could predict that those with a high school level of education watch an average  $5.30 + (-.17)(12)$  hours, or 3.26 hours per day.

The ANOVA table provides the results of the analysis of variance test. The table includes regression and residual sum of squares, as well as mean squares. To test the null hypothesis that  $r^2$  is zero, you will only need the statistic shown in the last column labeled “Sig.” This is the  $p$  value associated with the  $F$  ratio listed in the

Figure 10.15

The image shows the 'Linear Regression' dialog box in SPSS. On the left is a list of variables: ABANY, ABNOMORE, ABPOOR, AGE, AGEKDBRN, BIBLE, CHILDS, CHLDIDEL, CLASS, COEDUC, CONDOM, CONFED, CONPRESS, DEGREE, EDUC, EMAILHR, FEPOL, filter\_\$, and TVHOURS. The 'Dependent' variable is 'TVHOURS'. The 'Independent(s)' variable is 'EDUC'. The 'Method' is set to 'Enter'. There are buttons for 'Statistics...', 'Plots...', 'Save...', 'Options...', and 'Style...'. At the bottom are 'Reset', 'Paste', 'Cancel', and 'OK' buttons.

Figure 10.16

Model Summary

| Model | R                 | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1     | .177 <sup>a</sup> | .031     | .030              | 2.950                      |

a. Predictors: (Constant), Highest year of school completed

ANOVA<sup>a</sup>

| Model |            | Sum of Squares | df   | Mean Square | F      | Sig.              |
|-------|------------|----------------|------|-------------|--------|-------------------|
| 1     | Regression | 285.276        | 1    | 285.276     | 32.775 | .000 <sup>b</sup> |
|       | Residual   | 8799.836       | 1011 | 8.704       |        |                   |
|       | Total      | 9085.112       | 1012 |             |        |                   |

a. Dependent Variable: Hours per day watching tv

b. Predictors: (Constant), Highest year of school completed

Coefficients<sup>a</sup>

| Model |                                  | Unstandardized Coefficients |            | Standardized Coefficients | t      | Sig. |
|-------|----------------------------------|-----------------------------|------------|---------------------------|--------|------|
|       |                                  | B                           | Std. Error | Beta                      |        |      |
| 1     | (Constant)                       | 5.304                       | .418       |                           | 12.688 | .000 |
|       | Highest year of school completed | -.172                       | .030       | -.177                     | -5.725 | .000 |

a. Dependent Variable: Hours per day watching tv

column head “*F*.” The *F* statistic is 32.775, and its associated *p* value is .000. The model is significant. We therefore reject the null hypothesis at the .05 level.

**Demonstration 4: Producing a Multiple Regression Equation**

What other variables, in addition to education, affect the number of hours of television watched per day? One possible answer to this question is that age (AGE) has something to do with the television viewing hours per day. To answer this question, we will use SPSS to calculate a multiple regression equation and a multiple coefficient of determination. This procedure is similar to the one used to generate the bivariate regression equation. Click *Analyze*, *Regression*, then *Linear*. We place EDUC (number of years of education) and AGE (age in years) in the box for the independent variables and TVHOURS (the hours of television viewing per day) in the box for the dependent variable, and click *OK*. The output is presented in Figure 10.17.

Under the Model Summary, the multiple correlation coefficient labeled “*R*” is .269. This tells us that education and age are weakly associated with television viewing hours per day. The coefficient of determination is labeled “*R* square.” Its value is .073. An *R*<sup>2</sup> of .073 means that educational attainment and age jointly explain 7.3% of the variation in hours of television viewing per day. In addition, SPSS provides an “adjusted *R* square,” which is .071. The “adjusted *R* square”

**Figure 10.17**

Model Summary

| Model | R                 | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1     | .269 <sup>a</sup> | .073     | .071              | 2.892                      |

a. Predictors: (Constant), Age of respondent, Highest year of school completed

ANOVA<sup>a</sup>

| Model |            | Sum of Squares | df   | Mean Square | F      | Sig.              |
|-------|------------|----------------|------|-------------|--------|-------------------|
| 1     | Regression | 657.741        | 2    | 328.871     | 39.320 | .000 <sup>b</sup> |
|       | Residual   | 8405.862       | 1005 | 8.364       |        |                   |
|       | Total      | 9063.603       | 1007 |             |        |                   |

a. Dependent Variable: Hours per day watching tv

b. Predictors: (Constant), Age of respondent, Highest year of school completed

Coefficients<sup>a</sup>

| Model |                                  | Unstandardized Coefficients |            | Standardized Coefficients | t      | Sig. |
|-------|----------------------------------|-----------------------------|------------|---------------------------|--------|------|
|       |                                  | B                           | Std. Error | Beta                      |        |      |
| 1     | (Constant)                       | 3.472                       | .494       |                           | 7.034  | .000 |
|       | Highest year of school completed | -.162                       | .029       | -.167                     | -5.480 | .000 |
|       | Age of respondent                | .034                        | .005       | .204                      | 6.700  | .000 |

a. Dependent Variable: Hours per day watching tv

adjusts the  $R^2$  coefficient for the number of predictors in the equation. Generally, the adjusted  $R^2$  will be lower, relative to  $R^2$ , the larger the number of predictors.

The regression equation coefficients are listed in the Coefficients table. The regression equation coefficients are listed in the column headed “B.” The coefficient for EDUC is about  $-.162$ , and for AGE, it is  $.034$ . The intercept term, or  $a$ , identified in the “(Constant)” row, is  $3.472$ . Thus, we would predict that, holding age constant, every additional year of education decreases the number of television viewing hours per day by about 10 minutes ( $.162 \times 60$ ).

## SPSS PROBLEMS [GSS18SSDS-A]

- S1. Explore the relationship between the number of siblings a respondent has (SIBS) and his or her number of children (CHILDS).
  - a. Construct a scatterplot of these two variables in SPSS, and place the best-fit linear regression line on the scatterplot. Describe the relationship between the *number of siblings a respondent has* (independent variable) and the *number of his or her children* (dependent variable).
  - b. Calculate the regression equation predicting CHILDS with SIBS. What are the intercept and the slope? What are the coefficient of determination and the correlation coefficient?
  - c. What is the predicted number of children for someone with three siblings?
  - d. What is the predicted number of children for someone without any siblings?
- S2. Use the same variables as in Exercise 1, but do the analysis separately for men and women. Begin by locating the variable SEX. Click *Data, Split File*, and then select *Organize Output by Groups*. Insert SEX into the box and click *OK*. Now, SPSS will split your results by sex.
  - a. Calculate the regression equation for men and women. (*Note: You will need to scroll down through your output to find the results for men and women.*) How similar are they?
  - b. What is the predicted number of children for a man with six siblings? For a woman with the same number of siblings? Which group has the higher predicted number of children?
- S3. Use the same variables as in Exercise 1, but do the analysis separately for white and black respondents. Click *Data, Split File*, and then select *Organize Output by Groups*. Insert RACECEN1 into the box and click *OK*. SPSS will split your results by RACECEN1 (focusing your analysis only on the categories for whites and blacks).
  - a. Is there any difference between the regression equations for whites and blacks?
  - b. What is the predicted number for whites and blacks with the same number of siblings: one sibling, four siblings, and seven siblings?

- S4. Use the same variables as in Exercise 1, but do the analysis separately for married and divorced respondents. Begin by locating the variable MARITAL. Click *Data*, *Split File*, and then select *Organize Output by Groups*. Insert MARITAL into the box and click *OK*. SPSS will split your results by marital status.
- Is there any difference between the regression equations for married and divorced respondents?
  - What is the predicted number of children for married and divorced respondents with the following number of siblings: one sibling, four siblings, and seven siblings?
  - What differences, if any, do you find? Is the number of siblings a better predictor of number of children for married respondents or for women?
- S5. Investigate the relationship between the respondent's education (EDUC) and the education received by his or her father and mother (PAEDUC and MAEDUC, respectively).
- Calculate the correlation coefficient, the coefficient of determination, and the regression equation predicting the respondent's education with father's education only. Interpret your results.
  - Determine the multiple correlation coefficient, the multiple coefficient of determination, and the regression equation predicting the respondent's education with father's and mother's education. Interpret your results.
  - Did taking into account the respondent's mother's education improve our prediction? Discuss this on the basis of the results from S5b.
  - Using the regression equation from S5a, calculate the predicted number of years of education for a person with a father with 12 years of education. Then, repeat this procedure, adding in a mother's 12 years of education and using the regression equation from S5b.
  - Review the ANOVA results. Can you reject the null hypothesis that  $R^2 = 0$ ?

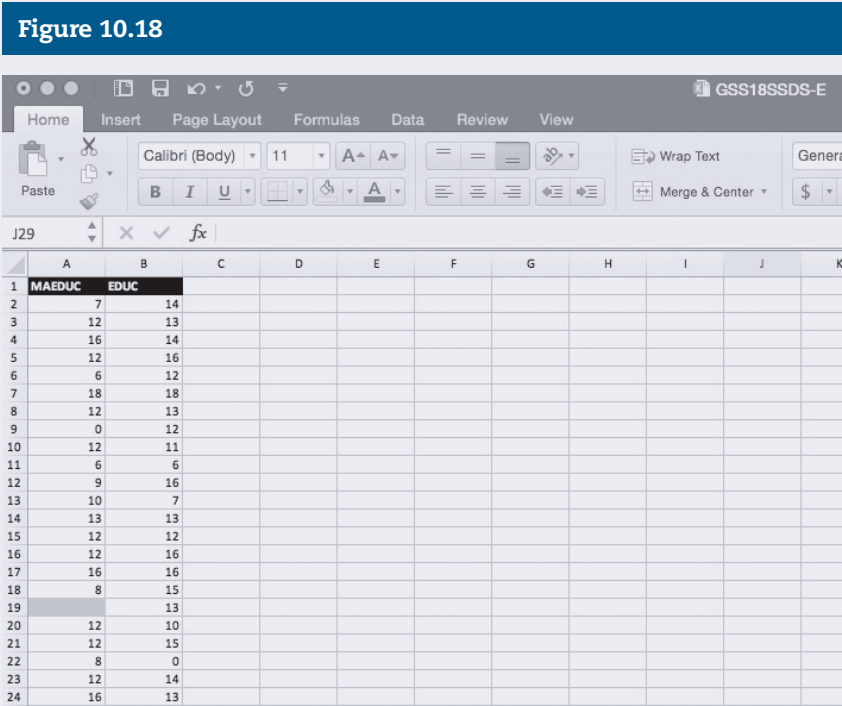
## EXCEL DEMONSTRATIONS [GSS18SSDS-E]

In this book, we do not demonstrate how to use Excel to produce a multiple regression prediction equation. While the steps are quite easy and involve using the "Data Analysis" function on Excel's *Data* tab, if you have any missing data in any of the variables you are working with, Excel will not let you proceed. There are ways around this ranging from deleting the cases with missing data, imputing missing data, and more. But, these techniques for dealing with missing data are beyond the scope of the basic statistics that we've covered in this text. Fortunately, missing

data is not an issue in Excel when our intent with regression analysis is limited to using the software to analyze bivariate data based on one interval-ratio independent variable and one interval-ratio dependent variable.

**Demonstration 1: Producing a Bivariate Linear Regression Equation**

In this demonstration, we will use Excel to produce a bivariate linear regression equation to examine the relationship between number of years of education one’s mother has completed (MAEDUC) and the number of years of education the respondents, themselves, have completed. We will treat MAEDUC as the independent variable and EDUC as the dependent variable. To begin our analysis, copy the MAEDUC and EDUC data from the protected Data View sheet and paste it into a new Excel sheet. You should paste the MAEDUC data in column A in the new Excel sheet and the EDUC data next to it in column B (see Figure 10.18). Placing MAEDUC and EDUC in the new sheet in this order will allow for us to more easily produce a scatter diagram of the data using Excel’s chart design—steps we will cover in Demonstration 3.





Excel is ready to calculate the value of the  $Y$  intercept ( $a$ ) and slope ( $b$ ) of our data. After we obtain these values, we will fill in our linear regression prediction equation:  $\hat{Y} = a + b(X)$ .

Let's begin with the  $Y$  intercept ( $a$ ). In cell D3, type " $Y$  intercept." You may wish to widen column D for organizational purposes. In cell E3, we will type `=INTERCEPT` and Excel's intercept function will appear. Notice how it reads "known\_ys, known xs" in parentheses after `INTERCEPT` (see Figure 10.19). Excel is instructing us to indicate the data for the dependent variable (the  $ys$ ) and the independent variable (the  $xs$ ).

We will complete the `=INTERCEPT` function by writing in cell E3 `=INTERCEPT(B2:B136, A2:A136)` and then *Enter* (see Figure 10.20). The value for our  $Y$  intercept will appear (9.57).

Now we will ask Excel to calculate the slope ( $b$ ) for our data. In cell D5, type "slope." In cell E5, we will type `=SLOPE` and Excel's slope function will appear. Notice how it reads "known\_ys, known xs" in parentheses after `SLOPE`

**Figure 10.19**

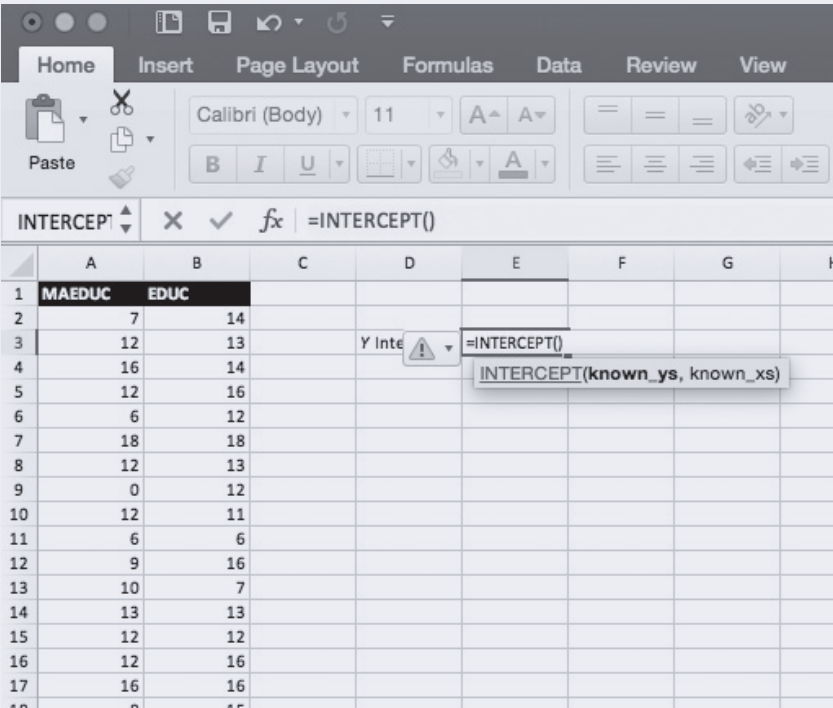
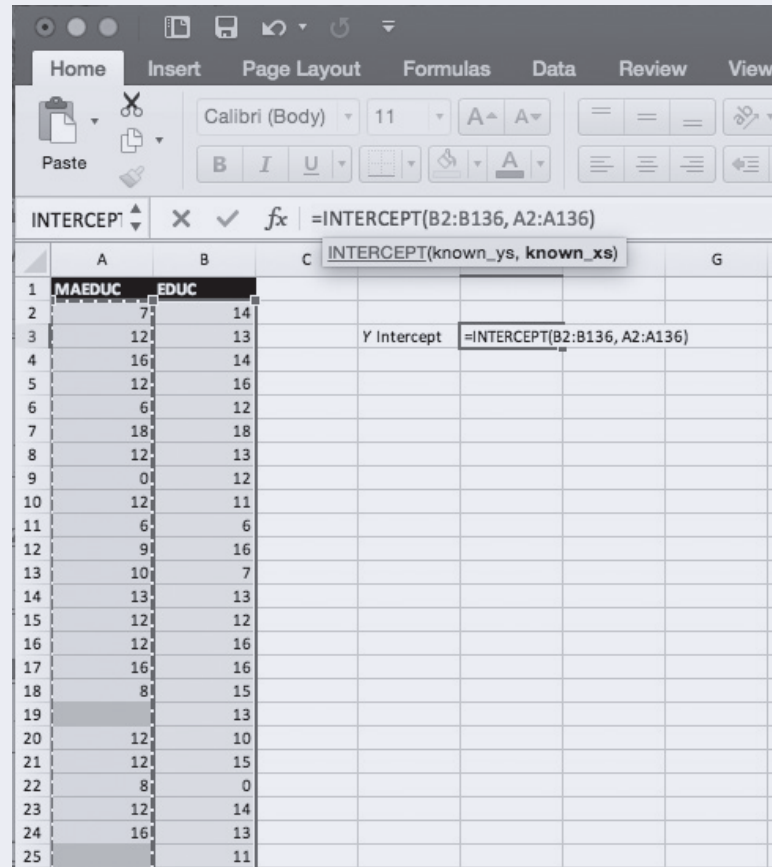


Figure 10.20



(see Figure 10.21). Excel is instructing us to indicate the data for the dependent variable (the *ys*) and the independent variable (the *xs*).

We will complete the `=SLOPE` function by writing in cell E5 `=SLOPE(B2:B136, A2:A136)` and then *Enter* (see Figure 10.22). The value for our slope will appear (.30).

We will plug in the value of both the *Y* intercept and slope into our regression prediction equation and come up with  $\hat{Y} = 9.57 + .30(X)$  (see Figure 10.23). Based on this equation, we can predict that if someone's mother completed zero years of formal education, we would predict this person's child to complete 9.57 years of education—the value of the *Y* intercept. For every year of education one's mother completed, we can predict their child will complete an additional .30 years of education—the value of the slope. As an example, if someone's mother completed 16 years of education (a bachelor's degree), we

Figure 10.21

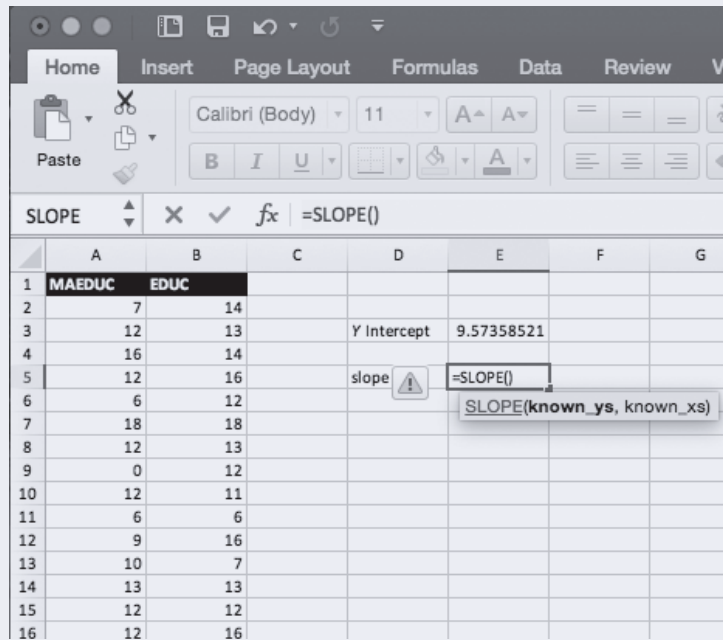
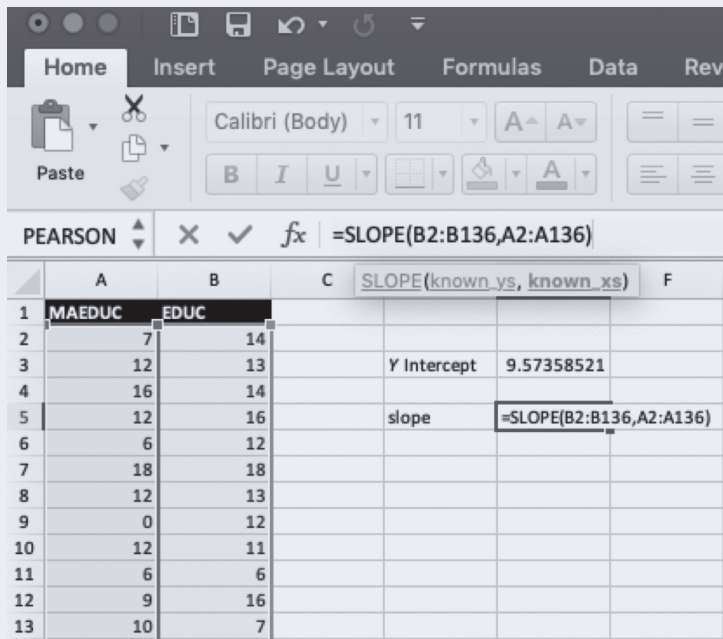
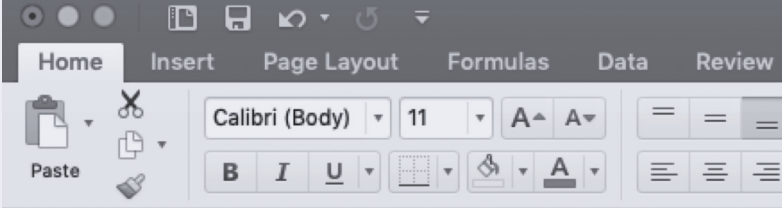


Figure 10.22



**Figure 10.23**



|    | A      | B    | C | D           | E          | F | G |
|----|--------|------|---|-------------|------------|---|---|
| 1  | MAEDUC | EDUC |   |             |            |   |   |
| 2  | 7      | 14   |   |             |            |   |   |
| 3  | 12     | 13   |   | Y Intercept | 9.57358521 |   |   |
| 4  | 16     | 14   |   |             |            |   |   |
| 5  | 12     | 16   |   | slope       | 0.30322112 |   |   |
| 6  | 6      | 12   |   |             |            |   |   |
| 7  | 18     | 18   |   |             |            |   |   |
| 8  | 12     | 13   |   |             |            |   |   |
| 9  | 0      | 12   |   |             |            |   |   |
| 10 | 12     | 11   |   |             |            |   |   |

would predict their child would complete 14.37 years of education ( $9.57 + 4.8 = 14.37$ ). Because the slope is positive, we could infer that there is a positive relationship between MAEDUC and EDUC.

### **Demonstration 2: Producing the Coefficient of Determination ( $r^2$ ) and the Correlation Coefficient ( $r$ )**

Continuing with the example from Demonstration 1, we will ask Excel to calculate the coefficient of determination ( $r^2$ ) for our data. In cell D7, type " $r^2$ ." In cell E7, we will type `=RSQ` and Excel's  $r^2$  function will appear. Notice how it reads "known\_ys, known\_xs" in parentheses after RSQ (see Figure 10.24). Excel is instructing us to indicate the data for the dependent variable (the ys) and the independent variable (the xs).

We will complete the `=RSQ` function by writing in cell E7 `=RSQ(B2:B136, A2:A136)` and then *Enter* (see Figure 10.25). The value for  $r^2$  will appear (.166).

Last, we will ask Excel to calculate Pearson's correlation coefficient ( $r$ ) for our data. In cell D9, type " $r$ ." In cell E9, we will type `=PEARSON` and Excel's Pearson function will appear. Notice how it reads "array1, array2" in parentheses after PEARSON (see Figure 10.26). Excel is instructing us to indicate the data for the two variables we are working with.

We will complete the `=PEARSON` function by writing in cell E9 `=PEARSON(B2:B136, A2:A136)` and then *Enter* (see Figure 10.27). The value for  $r$  will appear (.41).

Figure 10.24

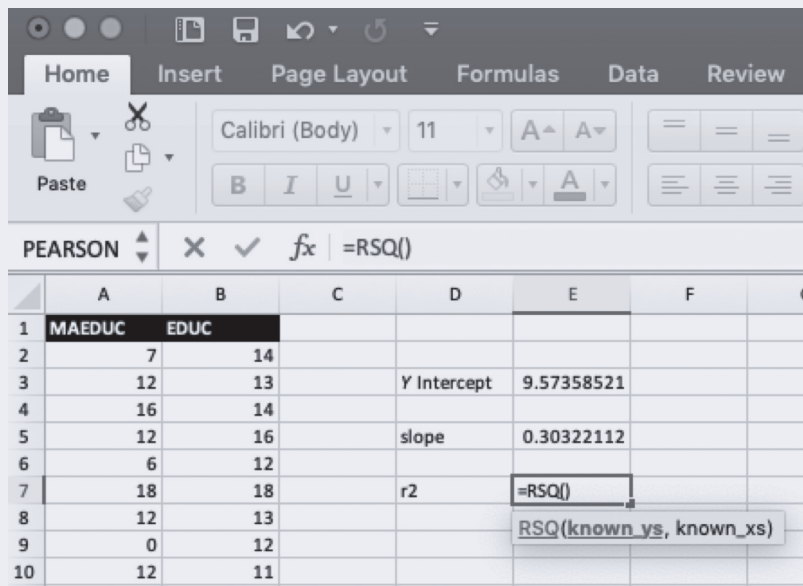


Figure 10.25

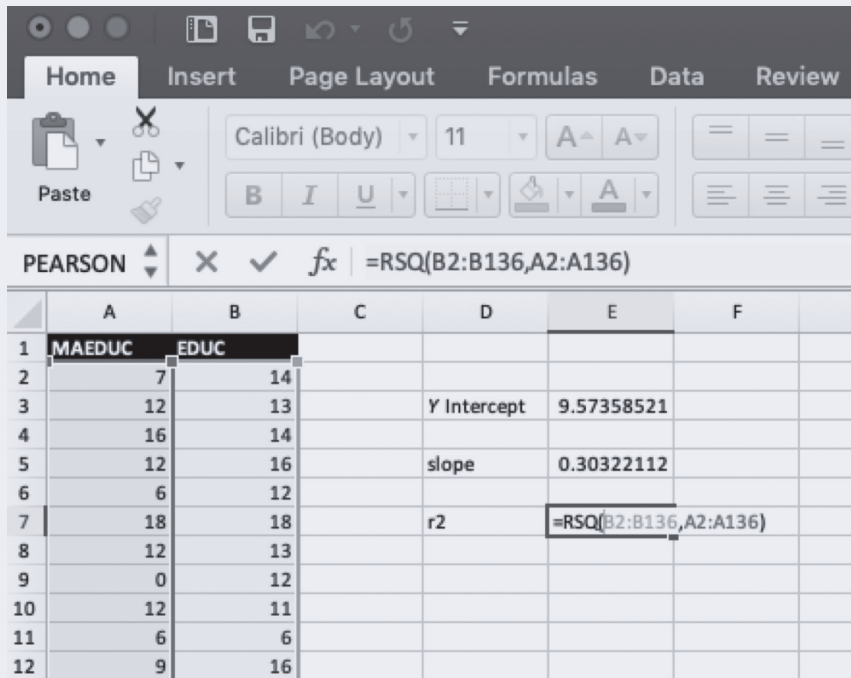


Figure 10.26

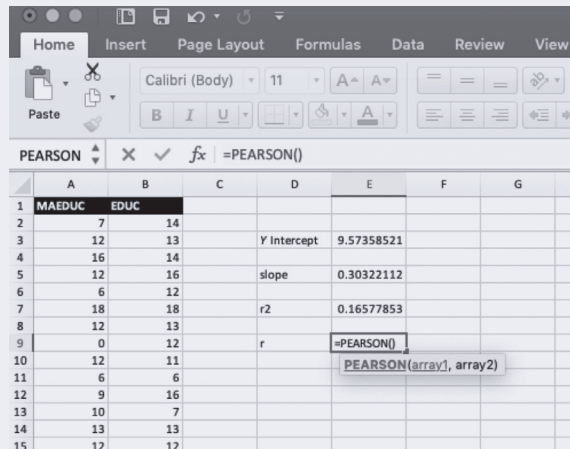
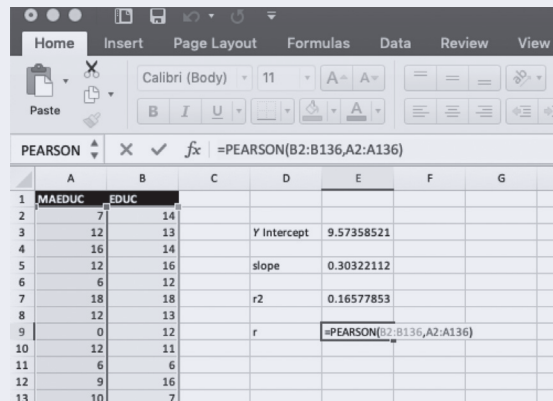
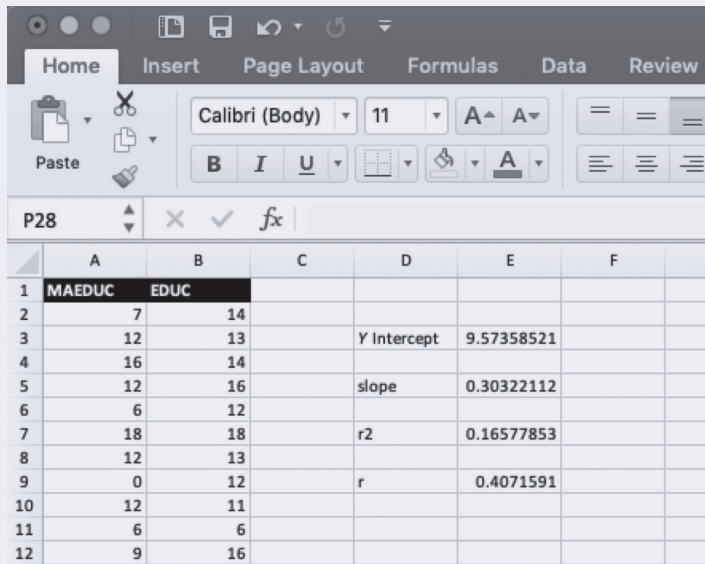


Figure 10.27



We can interpret the coefficient of determination ( $r^2$ ) and Pearson's correlation coefficient ( $r$ ) as follows: The  $r^2$  value is .166 or 16.6%, which indicates MAEDUC explains just over 16% of the variation of EDUC. By using MAEDUC to predict EDUC, we've reduced our prediction error by 16.6%. The Pearson's correlation coefficient ( $r$ ) is .41, which indicates a moderate to strong positive relationship between MAEDUC and EDUC. As MAEDUC increases, so does EDUC. The values of  $r^2$  and  $r$  are listed alongside the values for the  $Y$ -intercept and slope in Figure 10.28.

**Figure 10.28**



### **Demonstration 3: Producing a Scatter Diagram Using Excel's Chart Design**

We will produce a scatter diagram using the MAEDUC and EDUC variables we were working with in Demonstrations 1 and 2. Select all of the data in columns A and B from rows 2 to 136 (row 1 contains the labels MAEDUC and EDUC). On the “Insert” Excel tab, next to “Recommend Charts,” click on the dropdown image of the scatter diagram. There are many scatter diagrams to choose from. Let’s choose the first one under “Scatter” (see Figure 10.29).

Excel will create the scatter diagram treating the first column of data (MAEDUC) as the independent variable and thus placing it on the horizontal axis. Excel will treat the second column of data (EDUC) as the dependent variable and place it on the vertical axis. We’ve chosen to replace the standard “Chart Title” with “MAEDUC and EDUC.” To do this, simply click on “Chart Title” and revise the title as you see fit.

It is difficult to tell whether a relationship exists just by looking at points in the scatter diagram, so we will ask Excel to include the regression line. To add a regression line to our chart, double click on the chart to activate the *Chart Design* Excel tab. You will find an “Add Chart Element” drop-down option directly above column A (see Figure 10.30). Under the drop-down option for “Add Chart Element,” select *Trendline* → *Linear*. A regression line will appear on your scatter diagram.

Figure 10.29

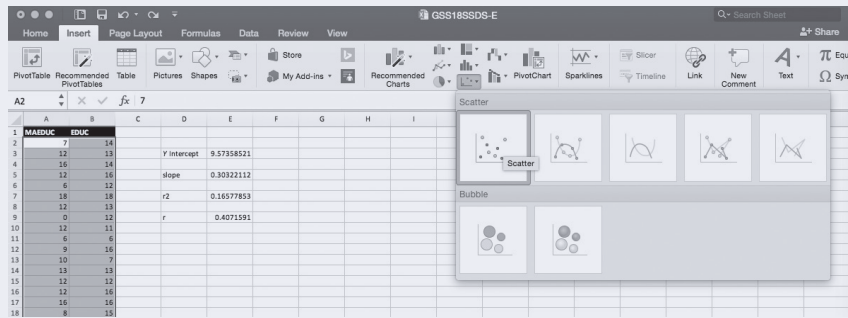
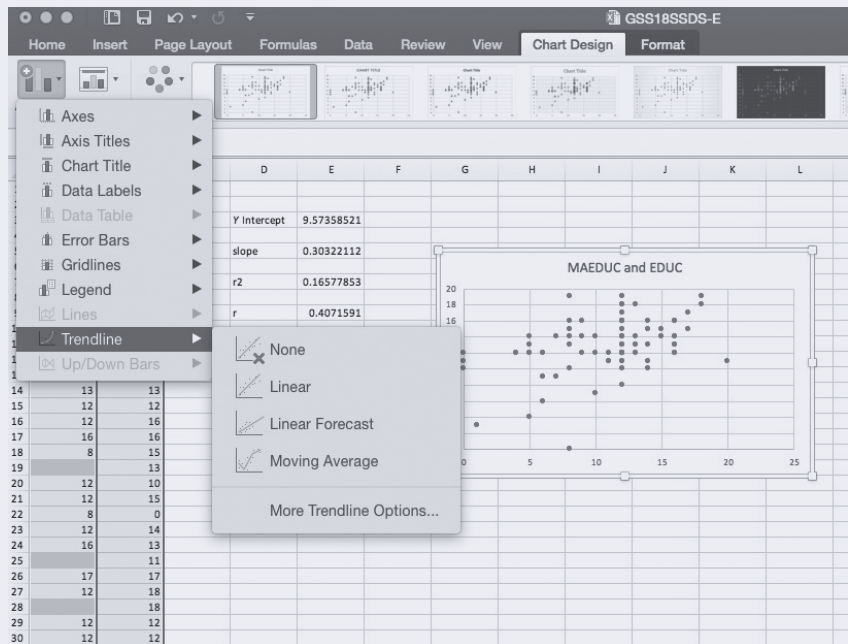


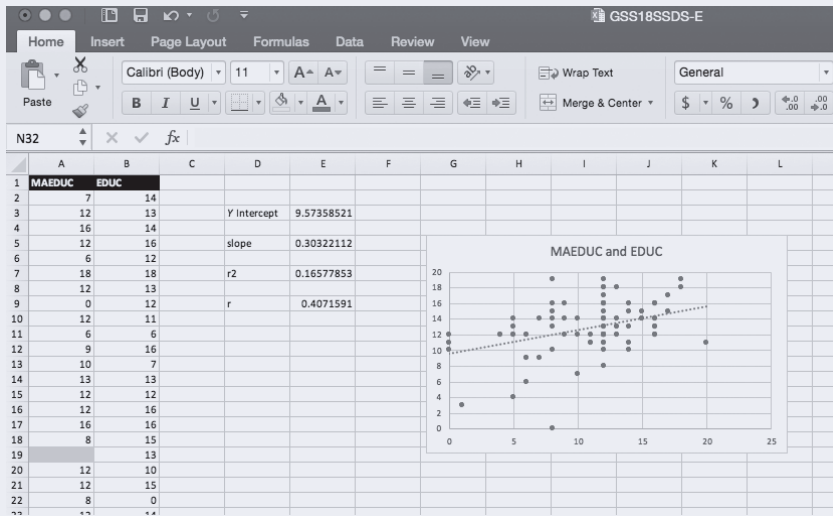
Figure 10.30



Since the regression line increases as mother's number of years of education increases, we observe a positive relationship between MAEDUC and EDUC (see Figure 10.31). However, because there is substantial scatter around the line (the points do not all perfectly align with the regression line), the predictive power of the model isn't deterministic (i.e., we are not working with a perfect linear relationship).



**Figure 10.31**



## EXCEL PROBLEMS [GSS18SSDS-E]

- E1. Examine the relationship between years of education (EDUC) and respondent's age when first child was born (AGEKDBRN). Treat EDUC as the independent variable and AGEKDBRN as the dependent variable.
  - a. Identify and interpret the value of the  $y$  intercept.
  - b. Identify and interpret the value of the slope.
  - c. Write out the regression equation predicting AGEKDBRN with EDUC.
  - d. Using the regression equation, how old is someone with 16 years of education predicted to be when their first child is born?
  - e. Create a scatter diagram of EDUC and AGEKDBRN. Impose a regression line on your visual.
- E2. Do folks who come from bigger families have more children? Examine the relationship between respondent's number of siblings (SIBS) and number of children (CHILDS) respondent has.
  - a. What is the independent variable? The dependent variable?
  - b. Write out the regression equation predicting CHILDS with SIBS.
  - c. Using the regression equation, if someone has three siblings, how many children are they predicted to have?
  - d. Identify and interpret the value of the coefficient of determination.

- e. Identify and interpret the value of the correlation coefficient.
  - f. Create a scatter diagram of SIBS and CHILDS. Impose a regression line on your visual.
- E3. Examine the relationship between ideal number of children (CHLDIDEL) and respondent's number of children (CHILDS). Treat CHLDIDEL as the independent variable and CHILDS as the dependent variable.
- a. Write out the regression equation predicting CHILDS with CHLDIDEL.
  - b. Using the regression equation, if someone's ideal number of children is five, how many children are they predicted to have?
  - c. Create a scatter diagram of CHLDIDEL and CHILDS. Impose a regression line on your visual.
  - d. Have Excel calculate the correlation coefficient for CHLDIDEL and CHILDS. Using the correlation coefficient, describe the relationship between the two variables.