7

LINEAR REGRESSION MODEL APPLICATIONS

Because this is the first application chapter, let me preview the basic structure and content that each application chapter has. I introduce the estimation technique by describing the situations and types of data for which the technique is appropriate and commonly used. I next summarize the properties that identify it as a GLM and discuss the meaning of the modeled outcome and its metric if they are different from the observed outcome. I then identify key diagnostic procedures, and I end the introduction with a note concerning the source of the data for the application examples. The remainder of the chapter is typically two start-to-finish application examples: (1) a single-moderator example and (2) a multiple-moderator and/or three-way interaction example. Each application example begins with a description of the dependent and independent variables and an overview of the results of diagnostic procedures and tests of the statistical significance of the interaction effects.

To briefly review, each GLM is built around a linear function of a set of predictors; that is, a sum of coefficients multiplied by predictors. The predictors may include interval variables, polynomials or other functions of an interval variable, categorical variables expressed as a set of binary indicators, and/or sets of predictors multiplied together (i.e., interaction terms). A specific GLM is distinguished by two key properties. The *link function* mathematically defines the transformation of the linear function of the predictors into the expected value of the outcome variable Y. Equivalently, we use the inverse of the link function, which describes the modeled outcome as a mathematical function of the observed outcome. The second distinguishing property is the *conditional distribution function*, which characterizes the probability distribution of the outcome conditional on the predictors.

OVERVIEW

Properties and Use of Linear Regression Model

Data and Circumstances When Commonly Used

Linear regression models (LRMs) are typically appropriate choices for analyzing interval or ratio outcome measures for which the mean of the outcome conditional on the predictors can be described by a linear function of the predictors. Note that the predictors can include categorical variables or a nonlinear function of a predictor, such as a polynomial or logarithmic expression, as well as interaction terms represented by the product of two or more predictors. But there are interval or ratio outcome measures for which an LRM can be a problematic choice. These include truncated or censored outcomes, badly skewed outcomes, mathematically bounded outcomes (e.g., a proportion), and count outcomes. A solution sometimes employed is to find a transformation of the outcome that makes an LRM appropriate (see Fox, 2008, chap. 4; Kaufman, 2013, chap. 3). The other, more common solution is to use a different member of the GLM family appropriate to the data situation.

Published Examples.

- Auspurg, Hinz, and Sauer (2017) used generalized least squares regression to analyze the fairness of baseline pay as portrayed in a series of vignettes to test theories of justice evaluation. Among other interactions, they tested whether the effect of the gendered pay ratio of respondents' job is moderated by whether respondents evaluate a man's base pay or a woman's base pay. They interpreted the interaction by describing how the focal variable's effect changed.
- In a two-stage least squares analysis of countries' tax revenues in the early modern era, Karaman and Pamuk (2013) estimated a three-way interaction between war pressure, urbanization, and regime representativeness. They created and interpreted a table of predicted tax revenues at selected values of urbanization and of real and hypothetical regime type.
- Cortes and Lincove (2016) studied the college admissions fit of student applicants (probability of a mismatch) using OLS regression and found an interaction between the applicants' high school class rank and their race/ethnicity. They discussed the interaction coefficients to interpret the interaction effect.
- Using OLS regression, Rambotti (2015) reanalyzed prior country-level research of the effect of income inequality and poverty on life expectancy to add a two-way interaction between income inequality and poverty. Rambotti interpreted the interaction by discussing predicted life expectancy plotted against inequality for low- and for high-poverty countries.

GLM Properties

The LRM is a GLM with an identity link function—Y equals the linear function of the predictors without a transformation of the prediction function—and a normal (Gaussian) conditional distribution function. In this instance, the modeled outcome is identical to the observed outcome. The LRM is a class of models with these characteristics, which are further differentiated by the assumptions they make about the error terms. OLS assumes that the error terms are normally distributed with equal variance (homoscedasticity) and zero covariance/uncorrelated errors (Greene, 2008, p. 111). Other models in this class relax the assumption of equal variance and/or zero covariance.[1] Common examples include WLS, which allows for a nonconstant error variance; time-series models, which usually specify autocorrelated errors; panel models; and least squares with robust standard errors, which may permit both correlated errors and unequal error variances. The empirical applications in this chapter use OLS in the two-moderator example and WLS in the one-moderator example.

Diagnostic Tests and Procedures

The diagnostic tests that I reviewed in Chapter 1 can all be used for LRMs. Plotting the residuals (typically studentized residuals) against the predicted outcome and against individual predictors provides an evaluation of overall model fit and possible model misspecification. These are essential for limiting the possible influence of a misspecified model on testing for and estimation of interaction effects. Similarly, such residual plots, leverage analysis, and influence statistics can help identify potential influential outliers to examine. There are also numerous tests of the assumptions of homoscedasticity and uncorrelated errors specific to OLS regression, panel models, and time-series analyses (for details, see Greene, 2008, chaps. 8, 9, and 19, respectively). Last, checking for collinearity among the predictors—apart from the necessary functional collinearity among interaction terms—is also standard.

Data Source for Examples

Both application examples use data from the 2010 GSS; the Stata data file (GSS_2010.dta) can be downloaded at www.icalcrlk.com, as can the Stata do-files used for the examples.

SINGLE-MODERATOR EXAMPLE

Data and Testing

The dependent variable is the frequency of sexual intercourse per month. The predictors consist of four interval measures (age in years, SES, number of children, and frequency of attending religious services) and two nominal measures (female-identified and never married status), each represented by a dummy variable. Both the frequency of sex and attendance at religious services were recoded from ordinal categories to a monthly frequency by annualizing the response category and dividing by 12.[2] The sample was restricted to respondents aged over 24; an additional 431 cases were excluded due to missing information on one or more of the variables.[3] Diagnostic testing indicated no outliers or influential cases but did suggest the presence of heteroscedasticity. A WLS regression was successful in correcting for heteroscedasticity and is used throughout the example. A sensitivity analysis showed no appreciable differences in the results reported for this example from using WLS versus OLS versus OLS with robust standard errors versus an OLS analysis of a variance-stabilizing transformation of the outcome (square root). Moreover, I treat the outcome measure of sexual intimacy as continuous to provide an example of a WLS analysis with heteroscedasticity, even though it has a limited number of observed values from the underlying continuous measure. Although doing so is somewhat contrived, the results are robust to alternative estimation choices and operationalizations of the outcome (negative binomial regression or ordinal logit).

The model includes an interaction between age and SES, conceptually justified by the expectation that the well-established negative effect of age on sexual activity would be diminished by the knowledge and resources available to those with higher SES. The WLS regression results report a t statistic for the interaction coefficient of 3.75 ($p < .001$). This indicates that the coefficient for the age-by-SES product term is statistically significant. Although the main effect coefficients for age and SES are each significant, this is not useful information in this case because 0 is not a valid value for either predictor. It is not meaningful to know that the effect of SES is significant for respondents who are 0 years old or that the effect of age is significant for respondents with *SES* = 0.

In the following sections, I cover the interpretation of the effect of age as moderated by SES and then the effect of SES as moderated by age, applying in turn the ICALC tools. The ICALC command lines are bolded in the Stata output for ease of identification.

The Effect of Age Moderated by SES

Setup With INTSPEC Tool

The first step is to use *intspec* to define the interaction details for the other tools:

```
. intspec focal(c.age) main( (c.age, name(Age) range(25(10)85)) ///
>         (c.ses, name(SES) range(17(10)97))) int2(c.age#c.ses) sumwgt(no)

Interaction Effects on Sexfrqmonth Specified as
-------------------------------------------------------

     Main effect terms: age  ses
     Two-way interaction terms: c.age#c.ses

  These will be treated as: Focal variable = age ("Age")
    moderated by interaction(s) with
        ses ("Ses")
```

Always check the output from *intspec* to verify that the interaction information is correct before proceeding to apply the other tools. The options/information specified are as follows:

- focal( ) declares *age* as the focal variable.
- main( ) contains the variable names of the main effect variables (*c.age* and *c.ses*) and their display names for the output, and range( ) lists the values used for display and/or calculation. The range specification for age—25(10)85—specifies that age's display values range from 25 to 85 in increments of 10. The range for SES is 10 to 97 in steps of 10.
- int2( ) specifies the variable name of the two-way interaction term *c.age #c.ses*.
- sumwgt(no) tells ICALC not to use the weights specified on the estimation command when calculating summary statistics. *This is crucial when your weights are analytic weights because the Stata margins command automatically uses the weights even when not listed in the margins command syntax unless you explicitly tell it not to do so.*

GFI Analysis

We start with the GFI tool to identify and extract the expression defining the effect of age and then to determine when, if at all, the sign of the effect of age changes. The option ndig(4) sets the display format for the coefficients to four digits after the decimal.

```
. gfi ,  ndig(4)

GFI Information from Interaction Specification of
Effect of Age on g(Sexfrqmonth) from Linear Regression
----------------------------------------------------------------------
Effect of Age =
   -0.2014 + 0.0013*Ses

 Sign Change Analysis of Effect of Age
 on g(Sexfrqmonth), Moderated by Ses (MV)

 --------------------------------------
             |           Age
 When        |  --------------------
 Ses=        |
 ------------+-------------------------
        17   | Neg b =    -0.1786
        27   | Neg b =    -0.1651
        37   | Neg b =    -0.1517
        47   | Neg b =    -0.1383
        57   | Neg b =    -0.1249
        67   | Neg b =    -0.1114
        77   | Neg b =    -0.0980
        87   | Neg b =    -0.0846
        97   | Neg b =    -0.0711
 ------------+-------------------------
 Sign Changes |       Never
 ------------+-------------------------
 % Positive  |         0.0
 --------------------------------------
```

The GFI expression tells us that the base effect of age is negative (−0.2014) and that *it becomes smaller in magnitude as SES increases*. Note that the italicized phrase would be incorrect if SES could have negative values. The sign change analysis table shows that within the range of sample values of SES, the effect of age is negative and never changes. The values of the moderated age effect change considerably. The largest magnitude effect (−0.1786) is more than 2.5 times the size of the smallest (−0.0711).

Significance Region Analyses: SIGREG and EFFDISP Tools

The SIGREG tool lets us explore whether the age effect changes in statistical significance as SES varies and the pattern and size of the changes in the age effect. I initially specify only one option: `ndig(3)` sets the number of digits for reporting the age effect to three.

Boundary Values for Significance of Effect of Age on **g(sexfrqmonth)** Moderated by SES
Critical value F = **3.848** set with p = **0.0500**

	When SES >=	Sig Changes	When SES >=	Sig Changes
Effect of age	116.745 (> max)	to Not Sig [-0.346]	254.014 (> max)	to Sig [0.035]

Note: Derivatives of Boundary Values in [ ]

Significance Region for Effect of Age (1 unit difference)
on g(sexfrqmonth) at Selected Values of SES

	At SES=								
Effect of	17	27	37	47	57	67	77	87	97
age	*-0.179**	*-0.165**	*-0.152**	*-0.138**	*-0.125**	*-0.111**	*-0.098**	*-0.085**	*-0.071**

Key: Plain font = Pos, Not Sig   **Bold font*** = Pos, Sig
Italic font = Neg, Not Sig   *Italic font** = Neg, Sig

Briefly, the boundary values report shows that the effect of age does not change significance within the sample range of SES values (17.1–19.2). There are changes in the age effect's significance that occur well beyond the maximum possible SES value of 97.2, when *SES* = 116.7 and *SES* = 254.0. The significance region table shows this invariant significance as well as the actual changes in the values of the age effect quite clearly. The effect of age on the frequency of intimacy is always negative. At the minimum SES (17), a 1-year increase in age would predict a 0.18 decline in the monthly frequency of intercourse. At the other end of the SES index, a 1-year increase in age would predict a 0.07 decrease in the frequency.

I think that the age effect's magnitude is better conveyed by reporting the age effect for a 10-year difference. And, given that the age effect is always negative and significant, limiting the display values for SES to every 20 units instead of every 10 units—the `range( )` suboption for SES on *intspec*—will produce a more succinct and readable table. Such results can be produced using the `effect( )` option as a calculator to get the focal variable's effect for any amount of change in the focal variable reported for each of the display values of the moderator. The syntax below accomplishes this recalculation by making the highlighted changes in the syntax for *intspec* and *sigreg*.

```
intspec focal(c.age) main( (c.age, name(Age) range(25(10)85)) ///
        (c.ses, name(SES) range(17(20)97))) int2(c.age#c.ses) ndig(0) ///
           sumwgt(n0)
sigreg, ndig(3) effect(b(10))
```

The results for the effect of age on frequency of intimacy for 10-year differences in age are as follows:

Significance Region for Effect of Age (10 unit difference)
on g(sexfrqmonth) at Selected Values of SES

	At SES=				
Effect of	17	37	57	77	97
age	*-1.786**	*-1.517**	*-1.249**	*-0.980**	*-0.711**

Key: Plain font = Pos, Not Sig   **Bold font*** = Pos, Sig
Italic font = Neg, Not Sig   *Italic font** = Neg, Sig

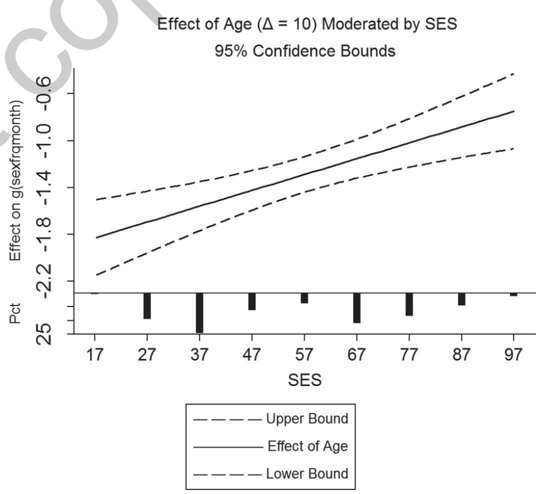
A 10-year difference in age predicts a decline of close to 2 times a month in the frequency of having sex for those with very low SES (17). For someone just above the mean of SES[4] (57), the decline is much less at 1.2 times fewer per month. And at high SES (97), a 10-year difference in age reduces the frequency of intercourse by less than once a month (0.7). Keeping in mind that the mean frequency of intimacy is 4 times per month, the reduction in the 10-year age effects with increasing SES appears substantial, as does the magnitude of the 10-year age effect at any level of SES. For example, the age effect at *SES* = 97 is a reduction equal to 18% of the mean, while the age effect when *SES* = 17 is 44% of the mean.

An alternative way to visualize how SES moderates the size, sign, and significance of the age effect is with a confidence bounds plot, the default plot type for an interval-by-interval interaction created with the EFFDISP tool. I add the `freq(tot)` option to display the relative frequency distribution of SES below the plot and use the `name( )` suboption within the `plot( )` option to store the plot as a memory graph with the specified name. This produces a report of the plot options in the results window and the graph shown in Figure 7.1.

```
. effdisp , effect(b(10)) plot(name(Age_by_SES_by_frq) freq(tot)) ndig(1)

Plot Options Specified

   name =  Age_by_SES_b_frq
   freq =  tot
   type =  cbound by default
```

This plot is a visual counterpart to the significance region table for 10-year age effects. The *solid black line* shows that the effect of age decreases in magnitude from about −1.8 to about −0.8 as SES rises from 17 to 97. The spike plot at the bottom of the graph shows a somewhat even distribution of cases between *SES* = 27 and *SES* = 77 with somewhat larger concentrations in the neighborhood of 37 and 67, places at

FIGURE 7.1 ⬢ EFFDISP PLOT FOR ONE MODERATOR



which the age effect is still fairly substantial. Because the confidence bounds (*dotted lines*) never contain the value 0, the age effect is always significant. Given that neither the sign of the age effect nor its statistical significance changes with SES, I find the effect display unnecessary. I would skip the effect display plot and instead present the significance region table or possibly an outcome display, as discussed next.

Outcome Displays: OUTDISP **Tool**

The OUTDISP tool produces tables or plots of the predicted outcome as it changes with the focal and moderating variables. Specifying the display values is especially consequential for creating tables because the predicted outcome is shown in the table only for the display values of the focal and moderating variables. For scatterplots and bar graphs, the moderators' display values similarly define and limit the calculation points (except that all the categories of a nominal variable are used to define calculation points). The focal variable's display values define axis labels, but the predicted outcome is calculated and shown across the full range of the focal variable's values. In general, pick the number and spacing of display values to capture important changes in the pattern of the relationship of the outcome with the interacting predictors.

As we just saw, the moderated age effect does not change sign or significance across the sample values of SES. Thus, using four display values across close to the full range of SES sample values (20–95 in increments of 25) works well to portray the relationship. This requires respecifying the range( ) suboption for SES on the *intspec* command and then running the *outdisp* command:

```
intspec focal(c.age) main( (c.age, name(Age) range(25(10)85)) ///
      (c.ses, name(SES) range(20(25)95))) int2(c.age#c.ses) ndig(0)
outdisp, outcome(atopt((means) _all)) plot(name(Age_by_SEI_Sex)) ///
   table(default)
```

The *outdisp* options specify the following:

- outcome(atopt((means) _all)) sets how the values of the predictors not part of the interaction are treated in calculating the predicted values of the monthly frequency of intercourse; that is, they are set to their means.[5]
- plot(name(SexFrq_by_Age_by_SES)) requests the creation of a plot that is stored during the duration of the Stata session with the specified name. Because the type( ) suboption is not listed, ICALC creates the default plot type, which is a scatterplot for an interval-by-interval interaction.
- table(default) specifies creation of a table of predicted values with the rows defined by the focal variable by default.

Reading down a column of the predicted values table indicates how the monthly frequency of having sex is predicted to change with age at the given level of SES. The pattern is easy to discern: The frequency of intercourse declines with age, and the amount of decline is less at higher levels of SES. A simple way to see this for a two-variable interaction—and explain it to an audience—is to compare for each column the change in the predicted outcome for the youngest age (25) with that for the oldest age (85). The reduction in the predicted number of times of having sex with age changes sharply with SES, as shown in Table 7.1:

- 10.473 for *SES* = 20
- 8.459 for *SES* = 45
- 6.444 for *SES* = 70
- 4.430 for *SES* = 95

The scatterplot in Figure 7.2 also portrays the pattern quite well. The lines in the scatterplot trace the predicted frequency of intercourse by age for each level of SES and make it quite obvious that all the slopes are negative. Similarly, how SES moderates the age effect is clear. The *solid line* for *SES* = 20 has the steepest slope. The slope is somewhat shallower when *SES* = 45 (*medium dashed line*), even shallower when *SES* = 70 (*large dashed line*), and shallower yet when *SES* = 95 (*small dashed line*).

Recap

The GFI and SIGREG results provided useful initial information about the moderation of the age effect by SES—namely, that the age effect is always negative and statistically significant across the range of SES values. The significance region table clearly summarized the relationship between age and frequency of intercourse and how that was moderated by SES. The confidence bounds plot produced by the EFFDISP tool also showed these patterns in a straightforward way. The application of the OUTDISP tool created both a table and a scatterplot of the predicted frequency of intimacy as it varies with age and how that relationship changes with SES. Both showed the pattern of the effect of age and its moderation by SES in a way that is easy to see and understand and, consequently, easy to present to an audience.

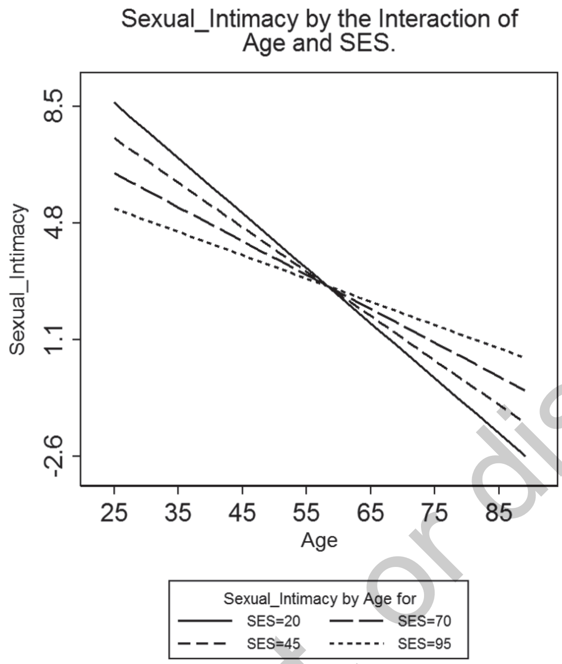
I suspect that many readers who are accustomed to seeing interaction effects interpreted using predicted values found the significance region table and confidence bounds plot less useful and harder to interpret, in part, because they are unfamiliar. If for nothing else, they are useful for identifying what the changes are in the pattern of the focal variable's effect and where they occur. This informs the choice of

TABLE 7.1 ⬢ PREDICTED VALUES TABLE FROM STATA OUTPUT

Predicted Value of sexfrqmonth by the Interaction of Age with SES.

	SES			
Age	20	45	70	95
25	**8.6077**	**7.4937**	**6.3797**	**5.2658**
35	**6.8622**	**6.0840**	**5.3057**	**4.5275**
45	**5.1167**	**4.6742**	**4.2317**	**3.7892**
55	**3.3712**	**3.2645**	**3.1577**	**3.0509**
65	**1.6257**	**1.8547**	**2.0837**	**2.3127**
75	**-0.1198**	**0.4449**	**1.0097**	**1.5744**
85	**-1.8653**	**-0.9648**	**-0.0643**	**0.8361**

FIGURE 7.2 ⬢ OUTDISP SCATTERPLOT FOR AGE MODERATED BY SES



display values for the moderator to show all the changes in the pattern. Hopefully, their utility will become more apparent when applied to more complex interactions or in situations where the moderated effect changes sign, as it does when we turn to interpreting the effect of SES moderated by age.

The Effect of SES Moderated by Age

One of the common errors in interpreting interaction effects that I discussed in Chapter 1 is to focus on only one of the predictors in the interaction and how it is moderated by the other. This provides an incomplete picture of the relationship between the outcome and the interacting predictors. As a case in point, we know how age affects the frequency of intercourse and how that is contingent on SES. But we know very little about how SES affects the outcome and how that is moderated by age. Thus, I reverse the roles of age and SES to interpret how SES as the focal variable is moderated by age.

Applying the ICALC Tools

The ICALC syntax that we use is nearly identical to what we just used for age as the focal variable. So I will walk through the example's Stata output and only note the changes in the command lines without a detailed explanation. Initially, I declare SES as the focal variable on the *intspec* command by changing the `focal(c.age)` option to `focal(c.ses)` and then use the same specifications for the *gfi* and *sigreg* commands.

```
intspec focal(c.ses) main( (c.age, name(Age) range(25(10)85)) ///
>         (c.ses, name(SES) range(17(10)97))) int2(c.age#c.ses) ndig(0) sumwgt(no)

Interaction Effects on Sexfrqmonth Specified as

     Main effect terms: ses  age
     Two-way interaction terms: c.age#c.ses

  These will be treated as: Focal variable = ses ("Ses")
    moderated by interaction(s) with
        age ("Age")

. gfi ,  ndig(4)

GFI Information from Interaction Specification of
Effect of Ses on g(Sexfrqmonth) from Linear Regression
----------------------------------------------------------------------

Effect of Ses =
   -0.0781 + 0.0013*Age

 Sign Change Analysis of Effect of Ses
 on g(Sexfrqmonth), Moderated by Age (MV)

 --------------------------------------
             |            Ses
 When        |  --------------------
 Age=        |
 ------------+-------------------------
        25   | Neg b =    -0.0446
        35   | Neg b =    -0.0311
        45   | Neg b =    -0.0177
        55   | Neg b =    -0.0043
        65   | Pos b =     0.0092
        75   | Pos b =     0.0226
        85   | Pos b =     0.0360
 ------------+-------------------------
 Sign Changes |  when MV= 58.17982
 ------------+-------------------------
 % Positive  |          29.6
 --------------------------------------
```

The algebraic expression for SES's effect tells us that its base effect on the frequency of having sex is negative (–0.0781) but becomes less negative (more positive) as age increases (0.0013). The sign change analysis indicates that the SES effect is initially negative but becomes smaller in magnitude as age increases and eventually turns positive. The row labeled "Sign Changes" reports that the sign change occurs when *Age* = 58.2; that is, for those 58 years and younger, there is a negative relationship between SES and frequency of intercourse; but for those older than 58, the relationship is positive. The last row in the table indicates that about one third of respondents have a positive effect of SES and two thirds have a negative effect.

Because the effect of SES changes sign, we want to know for what values of age the effect is significant. It could be that a range of both positive and negative effects is significant, that only positive effects are significant, or that only negative effects are significant. Knowing this is obviously consequential for how we understand the nature of the relationship between SES and the frequency of intimacy. Running the *sigreg* command produces the information we need.

```
Boundary Values for Significance of Effect of SES on g(sexfrqmonth) Moderated by Age
  Critical value F = 3.848 set with p = 0.0500
```

	When Age >=	Sig Changes	When Age >=	Sig Changes
Effect of ses	48.875	to Not Sig [-0.655]	69.251	to Sig [0.463]

```
Note: Derivatives of Boundary Values in [ ]

Significance Region for Effect of SES (1 unit difference)
on g(sexfrqmonth) at Selected Values of Age
```

	At Age=						
Effect of	25	35	45	55	65	75	85
ses	*-0.045**	*-0.031**	*-0.018**	*-0.004*	0.009	**0.023***	**0.036***

```
Key: Plain font  = Pos, Not Sig     Bold font*   = Pos, Sig
     Italic font = Neg, Not Sig     Italic font* = Neg, Sig
```

The boundary values analysis provides exact information on the change in significance of the effect of SES:

- When *Age* ≥ 48.875, the SES effect changes from significant to nonsignificant.
- When *Age* ≥ 69.251, the SES effect becomes significant again.

For this example, the significance region table works well to present how age moderates the SES effect because you can readily see the changing sign, magnitude, and statistical significance of the moderated effect. You can supplement discussion of the table with the exact age values at which the sign and significance of the effect change. But rather than show the SES effect as one-unit differences, I would present 1 standard deviation differences. And I would save the formatted table to Excel for presentation. You add two options to *sigreg* to do this:

```
sigreg , ndig(3) effect(b(sd)) save(Output\Table_7_2.xlsx table)
```

The `effect( )` option specifies calculating effects as 1 standard deviation changes in SES. In the `save( )` option, the keyword "table" saves the formatted table in the specified file and location.

Table 7.2 presents the formatted significance region table. This shows that the effect of SES on the frequency of having sex changes from negative and significant to positive and significant across the age range. For the youngest respondents (age 25), the frequency of intercourse is almost one time a month fewer for someone with 1 standard deviation higher SES. At age 45, the SES effect is a decline in frequency of about one third of a time per month, and the SES effect becomes not significant at age 48.875. The effect turns positive around age 58 but does not become significant until age is slightly more than 69 years. At age 75, a 1 standard deviation higher SES predicts someone having sex almost one-half time more per month, increasing to 0.7 times more for someone aged 85.

Applying the EFFDISP tool creates a confidence bounds plot by default that is also a very effective presentation choice in this context. Inside the `plot( )` option, the `name( )` suboption stores the plot as a named memory graph for the duration of your Stata session, and the `freq( )` option adds a spike plot of the frequency distribution of age to the plot shown in Figure 7.3. The *solid line* shows how the moderated effect of SES

TABLE 7.2 ⬢ EFFECT OF SES MODERATED BY AGE, FORMATTED TO HIGHLIGHT SIGN AND SIGNIFICANCE

Effect of SES (1 *SD* difference) on *g*(sexfrqmonth)							
	Age (years)						
Effect of SES	**25**	**35**	**45**	**55**	**65**	**75**	**85**
	–0.884*	***–0.618****	***–0.351****	*–0.085*	0.182	**0.448***	**0.715***
Key							
Plain font, no fill	Pos, not Sig						
Bold*, filled	Pos, Sig						
Italic, no fill	Neg, not Sig						
Bold italic*, filled	Neg, Sig						

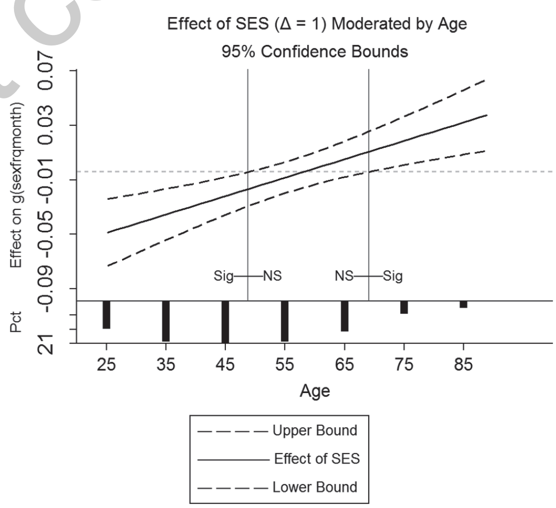
Note: SES = socioeconomic status; SD = standard deviation.

changes with age, while the two *dotted lines* represent the upper and lower confidence bounds for the SES effect. At a given age, when the *dotted lines* bracket the horizontal zero reference line, this indicates a nonsignificant effect.

```
effdisp, plot(name(SexFrq_SES_by_Age) freq(tot)) ndig(2)
```

The areas of significance and nonsignificance on the plot are shown by the two *vertical reference lines*. For ages below the first vertical line around age 49, SES has a negative and significant relationship with the frequency of having sex. For those between

FIGURE 7.3 ⬢ CONFIDENCE BOUNDS PLOT FROM EFFDISP



the two vertical lines—ages 49 through 69—SES does not have a significant effect on the frequency of having sex. And for those above age 69, the effect of SES turns positive and significant. Note from the spike plot that a plurality of cases have negative and significant effects of SES. The next largest group—ages 49 to 68—have no significant effect of SES, and the smallest group—over age 72—experience a positive relationship between SES and frequency of intercourse.

The OUTDISP tool creates tables and/or plots of the predicted outcome as it varies with the focal and moderating variables, which are the most commonly used displays for interpreting interaction effects. For this example, we want to pick display values to highlight how the SES effect varies across the range of ages from a positive to a negative effect of similar magnitude, and where the SES effect experiences notable transitions. Setting the age display to range from 26 through 81 in steps of 11 years will achieve this. The lines in a scatterplot (rows in the predicted values table) showing the relationship between SES and the frequency of sex contingent on age will be at or near where the SES effect

- has its minimum and maximum values (25, 85),
- changes from significant to nonsignificant (49),
- changes sign (58), and
- changes from nonsignificant back to significant (69).

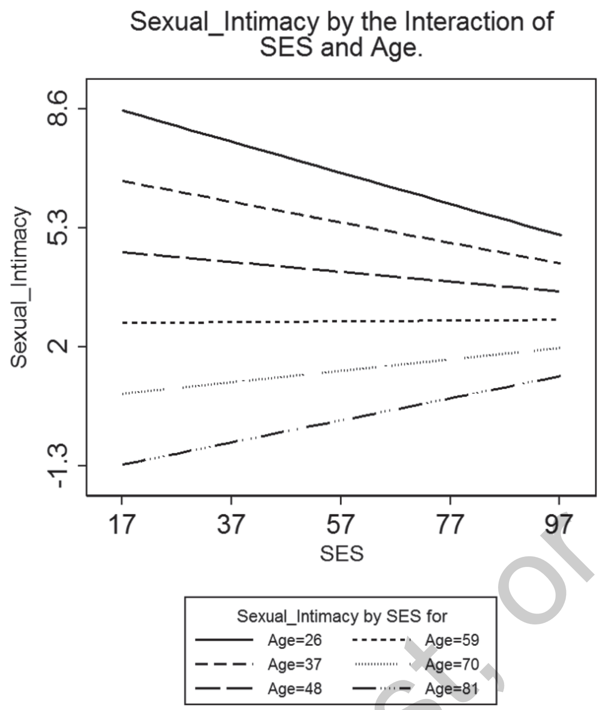
To accomplish this, we change the `range()` suboption for age in the *intspec* command to `range(26(11)81)` and then run the *outdisp* command with a `plot()` option to store the scatterplot to a memory graph, a `table()` option to create a table of predicted values, and the `outcome()` option to set the reference values for the other predictors, as shown in this excerpt from the Stata output.

```
. intspec focal(c.ses) main( (c.age, name(Age) range(26(11)81)) ///
   …

. outdisp, outcome(atopt((means) _all)) plot(name(SexFrq_by_SES_by_Age)) >
>       table(save(Output\Table_7_3.xlsx)
   …
```

The scatterplot in Figure 7.4 shows the progression of the SES effect on the frequency of intercourse from its most negative effect at age 26 (*solid line*) through its most positive effect at age 81 (*large-dash-and-dot line*). The third, fourth, and fifth lines from the top mark the transition from negative and barely significant (age 48), to barely positive and nonsignificant (age 58), to positive and barely significant (age 70). This scatterplot makes it easy to see the effect of SES and the frequency of having sex contingent on age. But while it is possible with some effort to tease out the effect of age contingent on SES from this plot, it is not easy to do so. I think this point is even truer for pulling out the effect of SES moderated by age from the scatterplot designed to highlight the effect of age moderated by SES in Figure 7.2. Presenting and interpreting both scatterplots would solve this problem.

A good alternative, at least for two-way interaction effects, is to report and interpret a table of predicted values from which you can equally well see and interpret both sides of the interaction effect. Look at Table 7.3, which presents the predicted values for frequency of having sex by age and SES. The rows are defined by SES and the

FIGURE 7.4 ● OUTDISP SCATTERPLOT OF SES MODERATED BY AGE



Note: SES = socioeconomic status.

TABLE 7.3 ● PREDICTED VALUES TABLE FROM OUTDISP, ONE MODERATOR

sexfrqmonth by the Interaction of SES With Age						
	Age					
SES	26	37	48	59	70	81
17	8.5628	6.5985	4.6341	2.6697	0.7054	–1.2590
37	7.6985	6.0296	4.3607	2.6917	1.0228	–0.6461
57	6.8342	5.4607	4.0872	2.7138	1.3403	–0.0331
77	5.9698	4.8918	3.8138	2.7358	1.6578	0.5798
97	5.1055	4.3229	3.5404	2.7578	1.9753	1.1927

columns by age, and the cell entries report the predicted frequency of intercourse for the combination of SES and age values. This table that ICALC saved to an Excel file is formatted with font sizes proportional to the cell values. Reading down a column shows the SES effect for the given value of age. In each of the first three columns, you can see that the predicted frequency of intercourse declines with SES as you read down the column—note that the font sizes are largest in the first row and then shrink as you read down the column.

You can also see that the rate of decline with SES diminishes as age increases across the columns. At age 26, the change in sexual activity from lowest to highest SES is a reduction of about 3.5 times per month (8.56 − 5.11 = 3.45), while at age 48, the reduction is only about once per month (4.63 − 3.54 = 1.09). Examining the next three columns, the SES effect has become positive, increasing across the columns. The increase in sexual activity with SES at age 59 is hardly different from 0 (2.76 − 2.67 = 0.09), but at age 81, the increase is about two thirds as large (1.19 − [−1.26] = 2.45) as the reduction in sexual activity with SES at age 25.

Examining the rows reveals the effect of age contingent on SES, which is even easier to discern because the frequency of having sex at every value of SES decreases from left to right—as do the font sizes—in each row as age increases. And the magnitude of the predicted drop in sexual activity declines from 9.82 at *SES* = 17 to about one third as large (3.91) at *SES* = 97. You can also see this by observing that the highest frequency of sexual activity in a row is for age 26 and that a row's maximum value steadily declines across the range of SES values. At the same time, the smallest frequency of sexual activity is always for age 81, and a row's minimum steadily increases across the range of SES values. Hence, the difference between the falling maximum and the rising minimum (i.e., the effect of age) steadily declines with SES.

Recap

The essence of how age moderates the relationship between SES and frequency of having sex is well documented by the GFI and SIGREG results: SES has a significant negative relationship with sexual activity for adults under age 49 but a positive relationship for those older than 69. Each of the summary displays of this relationship—the significance region table, the confidence bounds plot, the predicted values scatterplot, and the table of predicted values—made this overall pattern clear.

Summary and Recommendations

I explored a variety of techniques for probing this interval-by-interval interaction effect. The pattern of the relationship between monthly frequency of sexual activity and the interaction of age and SES is not complex. So you have many good choices to help you understand and then to explain to an audience the nature of the relationship. I think that presenting and interpreting any of the following could be effective:

1. The two confidence bounds plots for the effect of age on sexual activity contingent on SES and for the effect of SES on sexual activity contingent on age. This is my least preferred option in this example. These plots take more work to explain to most audiences, in part because they are unfamiliar and, likely, many readers will be confused at first because they expect to see predicted values rather than moderated effects. Thus, it may not be worth the effort to use them when other options show the pattern well and are easier to explain.
2. The two formatted significance region tables for the effect of age on sexual activity moderated by SES and for the effect of SES on sexual activity moderated by age. I would suggest reporting effects for 1 standard deviation

changes in the focal variable in both tables. In addition to creating a consistent presentation and discussion of effects, this also permits a comparison of the magnitude of the SES effect versus the age effect. I would also recommend removing the fill pattern highlighting. I think it is more distracting than helpful in this straightforward example. We will see its utility in later examples with multiple moderators or a three-way interaction.

3. The pair of scatterplots, one with plotted lines showing the predicted frequency of sex varying with age for selected values of SES and the other with plotted lines for the predicted frequency of intercourse changing with SES at selected values of age. These are visually appealing and easy to comprehend and interpret for an audience.
4. The table of predicted sexual activity as it changes simultaneously with age and SES. While the table has somewhat less immediate visual impact than the other options, it has the distinct advantage that you only need to present and discuss a single display of the relationship. Other than explaining that the font sizes are proportional to the predicted values, it requires little in the way of explanation or instruction about how to read the table and what it shows.

In the end, the choice is a matter of two factors—first, who your audience is and how accessible you think they would find the different modes of reporting the interactive relationship, and second, and equally important, what works best for you, what method you find the most intuitive and comprehensible. The better you can understand and relate to the technique for reporting the interaction effect, the better you will be able to use it to tell a story that others will understand.

TWO-MODERATOR EXAMPLE

Data and Testing

This analysis regresses the respondent's number of children on the interaction of family income and birth cohort, the interaction of family income and education, and also predictors of the number of siblings, religious intensity, and race. Children, income (in $10K), education, and siblings are interval measures. Birth cohort is a four-category nominal variable (Depression Era, WWII, Baby Boom, and Post-Boom[6]) represented by three dummy variables using Post-Boom as the reference category. Race is a three-category nominal variable (White, Black, and Other) represented by two dummy variables using White as the reference.

The sample is restricted to respondents aged 40 and over, to make it more likely that childbearing has been completed. An additional 218 cases were excluded due to missing information on one or more of the variables.[7] Diagnostic tests identified a small cluster of 15 outliers, but they were neither unusual (unrealistic) in their characteristics nor influential in affecting the estimation results and conclusions.[8] Statistical tests for heteroscedasticity indicated its possible presence, but diagnostic plots suggested a small degree of variation and were more consistent with nonlinearity in the effects of income and education, such as an interaction. A sensitivity analysis comparing the results for OLS versus OLS with robust standard errors demonstrated inconsequential differences, both overall and in terms of testing for the presence of the interaction effects.

This analysis demonstrates how to interpret interactions when there are two moderators of the same focal variable, family income by cohort and by education. Reversing the roles of the focal and moderator variables, it illustrates how to interpret an interaction between a multicategory focal variable and an interval moderator (cohort by family income), as well as a second example of an interval-by-interval interaction (education by family income). Conceptually, the interaction between economic conditions and birth cohort to predict fertility behavior has a long history of study, going back to the Easterlin (1961) hypothesis. The expectation is that the income effect should be larger (more negative) for more recent birth cohorts. An income-by-education interaction on fertility behaviors has also long been studied (e.g., Halli, 1990; Simon, 1975).

Including the income-by-education interaction in the model is supported by the *t* test of its coefficient ($t = 2.47$, $p = .14$) in the full model, which also includes the income-by-cohort interaction. The statistical grounds for including the income-by-cohort interaction terms in the model is to some degree a judgment call, depending on how you test the set of three interaction parameters (see Chapter 1). A global test of the simultaneous removal of the three parameters from the full model is not significant ($p = .166$), but the global test cannot take into account that we would expect all the three interaction parameters to be negative and hence that their sum should be less than 0.

You can do a one-tailed test of this under a null hypothesis that the sum of the three parameters is greater than or equal to 0. This results in a significant *t* statistic rejecting the null hypothesis ($t = -1.82$, $p = .035$). Alternatively, we could test the three parameters individually with a one-tailed test, adjusting the significance level for multiple testing, and conclude that the interaction term should be included if any of the three are negative and significant. This procedure also supports including the income-by-cohort interaction in the model because the Depression-Era-by-income coefficient is negative and significant ($t = -2.26$, $p = .012$, Sidak-adjusted $\alpha = .017$). Since the directional tests of the cohort-by-family-income interaction are significant, I include the family-income-by-cohort interaction in the model.

Strategy for Interpreting Two-Moderator Interaction Models

When you have multiple moderators of the same focal variable, you need to decide whether to interpret it one focal variable–moderator pair at a time or all of them simultaneously. In this instance, the first option is to interpret how the family income effect differs by cohort, with education set to reference values, and then how the income effect varies by education, with cohort limited to reference values. The second option is to interpret the effect of income as it changes across combinations of education's display values and cohort's display values. I prefer the second choice because it is a more holistic view. For LRMs and other linear link models, this choice is inconsequential because you will see exactly the same pattern for the income-by-education interaction at any value of cohort and analogously for income-by-cohort interaction.

Keep in mind that you should reverse the roles of the focal and moderator variables, which defines two focal-by-single-moderator interactions to explore—that is, (1) birth cohort's effect moderated by income and (2) education's effect moderated by income. I recommend interpreting these first. And then I would interpret the double-moderator component: how the family income effect is moderated by cohort and education. This will make an outcome display of the predicted number of children varying with income, cohort, and education simultaneously—which we will consider at the end of the section on interpreting family income moderated by cohort and education—easier to interpret because we will have an understanding of all the underlying patterns in hand.

The Effect of Birth Cohort Moderated by Family Income

INTSPEC **Setup and** GFI **Analysis**

These commands are shown in boldface at the top of the Stata output for the GFI analysis. I define birth cohort as the focal variable with the `focal(i.cohort)` option. The `main( )` option includes birth cohort and its moderator family income (*c.faminc10k*). Because *i.cohort* is a nominal factor variable, ICALC will automatically include all its categories as display/calculation values, so the `range( )` suboption would be ignored if it were listed. I set the display values for family income as $10K to $190K in $30K increments with the suboption `range(1(3)19)` because income is coded in units of $10K.

The *gfi* command has only the `ndig(3)` option to set the format for coefficients in the tables to three digits after the decimal. Given the scale of the coefficients for birth cohort and its interaction with family income, this provides three to four digits of information, which is what I prefer.

```
. intspec  focal(i.cohort)  ///
>         main((c.faminc10k , name(Family_Inc) range(1(3)19)) ///
>                 (i.cohort , name(Cohort))) ///
>         int2(c.faminc10k#i.cohort ) ndig(0)
   …
.
. gfi ,  ndig(3)

GFI Information from Interaction Specification of
Effect of Cohort on g(Childs) from Linear Regression
--------------------------------------------------------------------

Effect of BabyBoom =
   0.419 - 0.026*faminc10k

Effect of WWII =
   0.817 - 0.028*faminc10k

Effect of DeprEra =
   1.649 - 0.065*faminc10k
```

When Family_Inc=	Cohort BabyBoom		Cohort WWII		Cohort DeprEra	
1	Pos b =	0.393	Pos b =	0.789	Pos b =	1.585
4	Pos b =	0.316	Pos b =	0.705	Pos b =	1.391
7	Pos b =	0.239	Pos b =	0.620	Pos b =	1.197
10	Pos b =	0.162	Pos b =	0.535	Pos b =	1.004
13	Pos b =	0.085	Pos b =	0.451	Pos b =	0.810
16	Pos b =	0.008	Pos b =	0.366	Pos b =	0.616
19	Neg b =	-0.069	Pos b =	0.281	Pos b =	0.423
Sign Changes	when MV= 16.298		Never		Never	
% Positive	92.0		100.0		100.0	

Birth cohort has three effects in the regression analysis corresponding to the three included cohort indicators. Each effect represents the predicted difference in the number of children between an included cohort and the reference cohort (Post-Boom). Thus, the GFI and sign change analyses report information about each effect as it varies with family income. The GFI results indicate that each of the included cohorts has a positive main effect coefficient but a negative interaction effect coefficient. This means that each of these cohorts has a larger predicted number of children than the Post-Boom cohort when *Income* = 0 but the difference diminishes as income increases. The rate of decline is similar for the Baby Boom and WWII cohorts and less than half the rate for the Depression Era cohort. The sign change analysis tells us that only the Baby Boom effect turns negative (fewer predicted children than the Post-Boom cohort) and that this occurs at a fairly high income level ($162,980).

Significance Region Analyses: SIGREG and EFFDISP Tools

Because the differences among cohorts in the predicted number of children declines with family income, this raises the question of whether the differences remain significant, especially since they change at different rates. I start exploring this using the *sigreg* command with minimal options (shown in bold at the top of the output), specifying the significance level to use (.05) and the number of digits to report in tables. An important point to keep in mind is that the details of these results are contingent on the choice of the reference category for birth cohort. For instance, if we used the Depression Era cohort for the reference, almost all of the moderated effects for the three included cohorts would be negative rather than positive. But the results and hence the stories we tell would be consistent in the meaning of the overall cohort effect on the number of children.

Boundary Values for Significance of Effect of Cohort on **g(childs)** Moderated by Family Inc
Critical value F = **3.850** set with p = **0.0500**

Effect of Cohort	**When Family_Inc >=**	**Sig Changes**	**When Family_Inc >=**	**Sig Changes**
BabyBoom	5.919	**to Not Sig** [-0.457]	-2.338 (< min)	**to Sig** [0.142]
WWII	11.575	**to Not Sig** [-1.070]	-11.095 (< min)	**to Sig** [0.202]
DeprEra	15.213	**to Not Sig** [-1.356]	376.357 (> max)	**to Sig** [0.001]

Note: Derivatives of Boundary Values in [ ]

Significance Region for Effect of Cohort (1 unit difference)
on g(childs) at Selected Values of Family_Inc

	At Family Inc=						
Effect of	1	4	7	10	13	16	19
BabyBoom	**0.393***	**0.316***	0.239	0.162	0.085	0.008	*-0.069*
WWII	**0.789***	**0.705***	**0.620***	**0.535***	0.451	0.366	0.281
DeprEra	**1.585***	**1.391***	**1.197***	**1.004***	**0.810***	0.616	0.423

Key: Plain font = Pos, Not Sig **Bold font*** = Pos, Sig
Italic font = Neg, Not Sig *Italic font** = Neg, Sig

The boundary values analysis pinpoints the income level at which each of the cohort effects changes from significant to not significant. This occurs at $59,190 (5.919 × $10,000) for the Baby Boom cohort, about $116K for the WWII cohort, and about $152K for the Depression Era cohort. You can see this visually in the formatted significance region table—note where the cell entries change from having a * symbol to not—which also shows the Baby Boom effect turning negative but remaining nonsignificant (italicized font).

The effect values show the differences in the predicted number of children between each cohort compared with the Post-Boom cohort; positive values indicate that the

Post-Boom cohort has a smaller predicted number than the comparison cohort. The Post-Boom cohort has the lowest predicted number of children, with the exception of the Baby Boom cohort at high income levels, but this disparity is not significant. The other two cohorts always have a predicted number of children greater than the Baby Boom and Post-Boom cohorts. And the Depression Era cohort has a higher predicted number of children than the WWII cohort except at very high incomes.

Another good presentation option is a confidence bounds plot for each of the included cohort effects. Because the default plot is needed, I only specify on the *effdisp* command the option ndig(1) to format the *y*-axis labels to one decimal place:

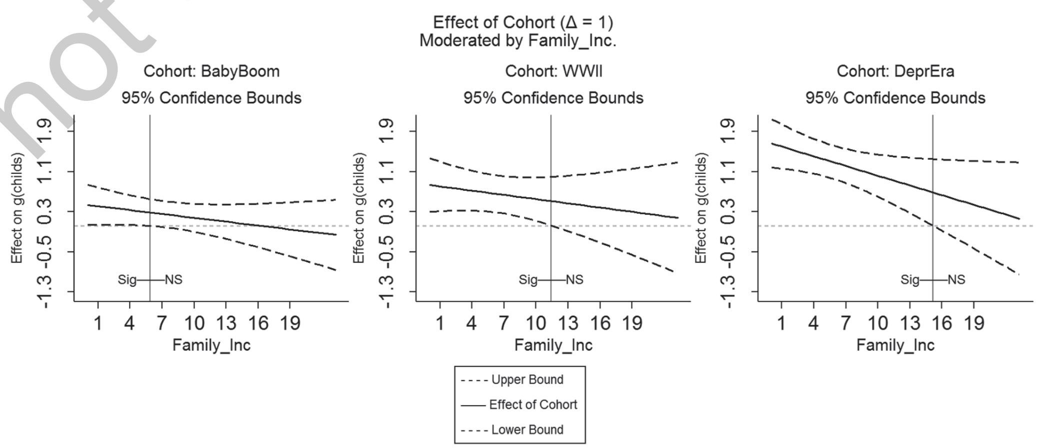
```
effdisp, ndig(1)
```

Figure 7.5 presents the three confidence bounds plots. The *horizontal thin reference line* separates the negative from the positive effects and reaffirms that only the Baby Boom cohort's effect turns negative. The *vertical reference lines* mark the change from significant to not significant for each cohort, as well as documenting the different income levels at which this occurs for each cohort.

Outcome Displays: OUTDISP **Tool**

For a multicategory focal variable, the results in outcome displays are not contingent on the choice of the base reference category because the outcome's predicted values are displayed for all the categories. Thus, they are typically a better option to convey the nature of the interaction effect in this situation than significance region tables or effect displays. I request the default plot from the *outdisp* command by specifying the options with plot(def) and use the moderator to define rows in the predicted values table with tab(row(mod)). The default plot for a categorical focal variable such as birth cohort is a bar chart for each of the display values of the moderator. I use the same *intspec* command as before, and the *outdisp* command produces the following output and bar charts.

FIGURE 7.5 ⬢ EFFDISP FOR MULTIPLE-CATEGORY FOCAL VARIABLE

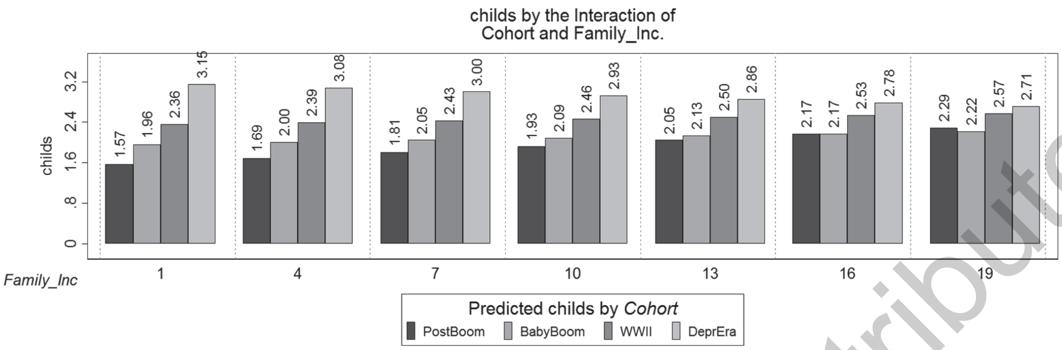


```
. intspec  focal(i.cohort)  ///
>         main((c.faminc10k , name(Family_Inc) range(1(3)19)) ///
>                 (i.cohort , name(Cohort) range(1/4))  ) ///
>         int2(c.faminc10k#i.cohort ) ndig(0)
    …
.
. outdisp, out(atopt((means) _all)) plot(def) table(row(mod)) ndig(2)

Outcome Options Specified or Default

   metric =  obs
   atopt =  (asobs) _all

Table Options Specified or Default

   row =  mod

Plot Options Specified or Default

   type =  bar

Predicted Value of Childs by the Interaction of Cohort with Family_Inc.

--------------------------------------------------------------
                   |                   Cohort
        Family_Inc |  PostBoom   BabyBoom     WWII    DeprEra
-------------------+------------------------------------------
                 1 |      1.57       1.96       2.36       3.15
                 4 |      1.69       2.00       2.39       3.08
                 7 |      1.81       2.05       2.43       3.00
                10 |      1.93       2.09       2.46       2.93
                13 |      2.05       2.13       2.50       2.86
                16 |      2.17       2.17       2.53       2.78
                19 |      2.29       2.22       2.57       2.71
---------------------------------------------------------------
```

Comparing the rows in the predicted values table reveals how the predicted differences among cohorts in number of children change across family income levels. The first five rows show the predicted number of children increasing from the youngest cohort (Post-Boom) to the oldest cohort (Depression Era). Doing some calculations in your head (or otherwise) shows diminishing differences among the cohorts at higher income levels. Moreover, the pattern of cohort differences alters in the last two rows while generally continuing to equalize. You can also contrast the patterns in the columns to see how the effect of family income is contingent on birth cohort. The Post-Boom cohort exhibits a steady increase in the predicted number of children with rising family income, the Depression Era cohort shows a smaller steady decline with income, and the other two cohorts show increases that are less in magnitude.

The bar charts in Figure 7.6 reveal these patterns without you having to do calculations in your head. The height of the bars represents the predicted number of children in a birth cohort, and the bars are labeled with the predicted value. When family income is $130K or less (the first five bar charts), there is a stair-step pattern in which the number of children increases from the most recent cohort to the oldest cohort. But the size of the steps diminishes as income rises. Beyond $130K, the cohort differences continue to level off, and the relative order of the number of children across cohorts changes.

FIGURE 7.6 ⬢ OUTDISP FOR MULTIPLE-CATEGORY FOCAL VARIABLE



Next, compare the height of the bar for each cohort individually across income levels. The Post-Boom cohort exhibits a steady increase in the predicted number of children with rising family income, the Depression Era cohort shows a steady decline with income, and the other two cohorts show increases that are much smaller in magnitude. This provides a first take on how family income is differentially related to the number of children. But the income effect is also moderated by education, so these values represent the income effect at education's reference value (its mean). In the next section, I probe how education's effect on the number of children is moderated by family income, and in the process, we learn more about the children–income relationship.

The Effect of Education Moderated by Family Income

INTSPEC **Setup and** GFI **Analysis**

The *intspec* command has the same structure as for the cohort-by-income interaction just analyzed, with the substitution of education (*c.ed*) and its information as the focal variable, as shown in the Stata output. The *gfi* command now reports four digits for the coefficient values in the tables.

```
. intspec  focal(c.ed)  ///
>         main((c.faminc10k , name(Family_Inc) range(1(3)19)) ///
>                 (c.educ , name(Education) range(0(4)20)) ) ///
>         int2( c.faminc10k#c.educ ) ndig(0)
   …
. gfi ,  ndig(4)

GFI Information from Interaction Specification of
Effect of Education on g(Childs) from Linear Regression
--------------------------------------------------------------------

Effect of Educ =
   -0.1836 + 0.0065*Faminc10k
 Sign Change Analysis of Effect of Education
 on g(Childs), Moderated by Family_Inc (MV)
```

```
-------------------------------------
             |      Education
When         |  --------------------
Family_Inc=  |
-------------+-----------------------
        1    | Neg b =    -0.1771
        4    | Neg b =    -0.1576
        7    | Neg b =    -0.1381
       10    | Neg b =    -0.1186
       13    | Neg b =    -0.0990
       16    | Neg b =    -0.0795
       19    | Neg b =    -0.0600
-------------+-----------------------
Sign Changes |        Never
-------------+-----------------------
% Positive   |          0.0
-------------------------------------
```

The GFI expression tells us that the moderated education effect starts as negative when family income is 0 and is predicted to decline by nearly 2/10 of a child for a 1-year difference in education. But the effect becomes less negative as family income rises—the coefficient value increases by 0.0065 with a $10K increase in family income. The sign change analysis shows that the education effect remains negative across the range of income values.

Significance Region Analyses: SIGREG Tool

To determine if the education effect remains significant, I use the *sigreg* command, with the number of digits for reporting set to four. The boundary values analysis indicates that the education effect on number of children is no longer significant once family income is greater than $170,529. The significance region table also shows this, as well as how the magnitude of the education effect is changing. By the income level at which it loses significance, the education effect is about one third of what it is at *Income* = 0. A confidence bounds plot could also be used to show this—*effdisp* with no options—but I prefer the significance region table when the results are simple and straightforward.

Boundary Values for Significance of Effect of Education on **g(childs)** Moderated by Family Inc
Critical value F = **3.850** set with p = **0.0500**

	When Family_Inc >=	**Sig Changes**	**When Family_Inc >=**	**Sig Changes**
Effect of educ	17.0529	to Not Sig [-1.2052]	159.7610 (> max)	to Sig [0.0087]

Note: Derivatives of Boundary Values in [ ]

Significance Region for Effect of Education (1 unit difference)
on g(childs) at Selected Values of Family_Inc

	At Family_Inc=						
Effect of	1	4	7	10	13	16	19
educ	***-0.1771****	***-0.1576****	***-0.1381****	***-0.1186****	***-0.0990****	***-0.0795****	*-0.0600*

Key: Plain font = Pos, Not Sig   **Bold font*** = Pos, Sig
Italic font = Neg, Not Sig   *Italic font** = Neg, Sig

Outcome Displays: OUTDISP Tool

I use the *outdisp* command to produce the table below and the plot in Figure 7.7 to portray how the predicted number of children changes with education contingent on family income. As shown at the top of the Stata output, I specify three options:

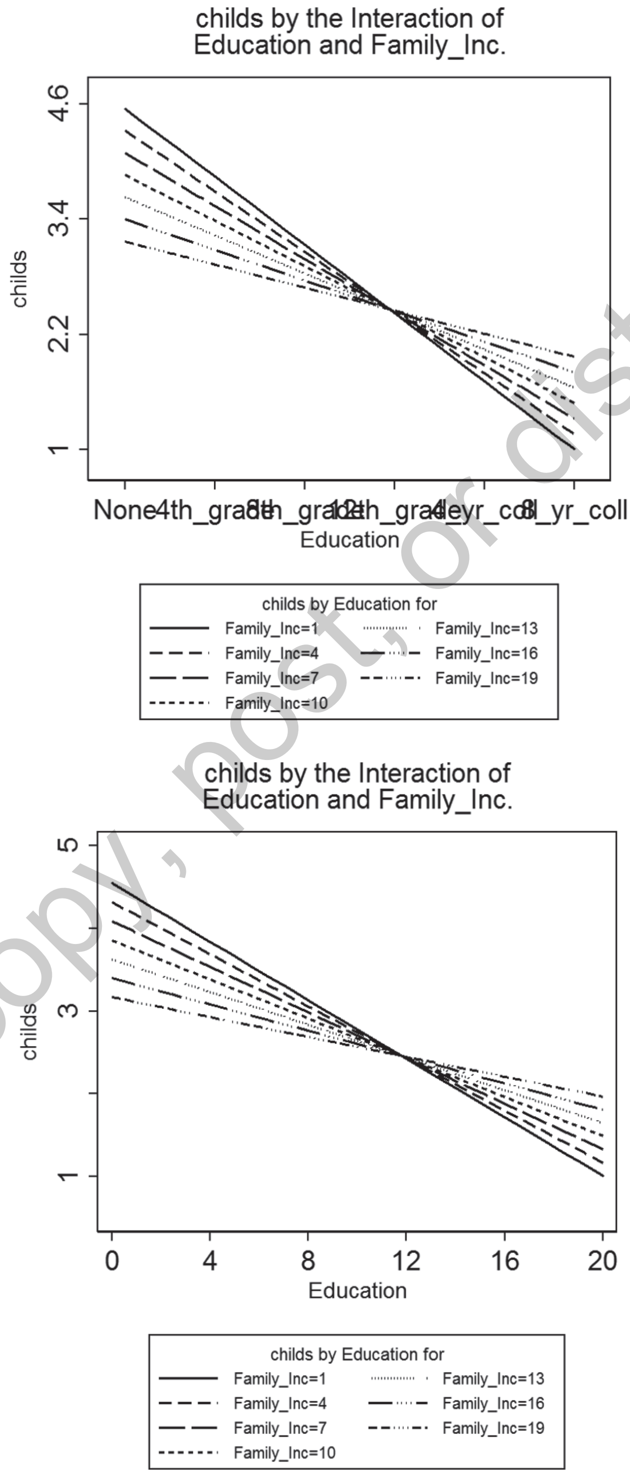
(1) out( ) to set the reference values for the other predictors in the model to their means, (2) plot( ) to produce the default scatterplot for interval variables stored as a named memory graph, and (3) tab( ) to create a predicted values table with the moderator (income) to define the rows.

```
. intspec  focal(c.ed)  ///
>          main((c.faminc10k , name(Family_Inc) range(1(3)19)) ///
>                  (c.educ , name(Education) range(0(4)20)) ) ///
>          int2( c.faminc10k#c.educ ) ndig(0)
    …
.
. outdisp, out(atopt((means) _all)) plot(name(Plot1)) tab(row(mod))

Outcome Options Specified or Default

   metric =  obs
   atopt =   (means) _all

Table Options Specified or Default

   row =  mod

Plot Options Specified or Default

   type =  scat
   name =  plot1


Predicted Value of Childs by the Interaction of Education with Family_Inc.


-----------------------------------------------------------------------------------
                |                              Education
     Family_Inc |          0           4           8          12          16          20
----------------+------------------------------------------------------------------
              1 |     4.5489      3.8404      3.1320      2.4235      1.7150      1.0065
              4 |     4.3189      3.6885      3.0581      2.4277      1.7973      1.1669
              7 |     4.0888      3.5365      2.9842      2.4319      1.8796      1.3272
             10 |     3.8587      3.3845      2.9103      2.4361      1.9618      1.4876
             13 |     3.6287      3.2325      2.8364      2.4403      2.0441      1.6480
             16 |     3.3986      3.0806      2.7625      2.4445      2.1264      1.8084
             19 |     3.1686      2.9286      2.6886      2.4487      2.2087      1.9687
-----------------------------------------------------------------------------------
```

Each row in the table shows that the predicted number of children decreases across education levels at a rate that diminishes with income level. For example, between 0 and 20 years of education, the number of children is predicted to drop by 3.54 children when income equals \$10K, by 2.37 when income equals \$100K, and by 1.20 for an income of \$190K. Comparing the columns reveals the relationship between number of children and family income as the relationship varies by education, but keep in mind this is at the reference values for birth cohort. Nonetheless, we see that the income effect changes from a negative one for low levels of education to a positive one at higher levels of education.

Presenting a predicted values plot is often a more convenient choice because it obviates the need for you to write out, as I did, examples of the magnitude of the numeric changes, to include such change calculations in the table, or to leave it to the reader to do head calculations. The upper panel of Figure 7.7 shows the default scatterplot created by the *outdisp* command. The lower panel is the same scatterplot with the

FIGURE 7.7 ⬢ OUTDISP SCATTERPLOT FOR ONE MODERATOR, DEFAULT AND REVISED



axis labeling cleaned up using the pltopts( ) option, as described in the concluding "Special Topics" section. Each line represents the children–education relationship for a given level of family income. The *solid line* shows that the predicted number of children drops most quickly across education levels when family income is near its minimum. The remaining *dashed lines* show that the children–education relationship becomes less and less steep as income rises. And from the boundary values analysis, we know that the slopes of the lines for the highest two income levels are the shallowest and not significantly different from 0.

In the end, I think that there are three good options for reporting how the effect of education on number of children is moderated by family income (four if you count a confidence bounds plot). The significance region table gives a simple and compact presentation of the information. A predicted values table with an added column of changes in the predicted number of children across the range of education for each level of family income would also be very effective and accessible (see the later discussion of Table 7.5). For a graphical presentation, I would recommend the scatterplot of predicted values both for the visual appeal of the plotted lines and for ease of comprehension and interpretation.

The Effect of Family Income Moderated by Birth Cohort and Education

Analyzing how income moderates the cohort effect and how it moderates the education effect gave insight into how the income effect is moderated by cohort and education separately. I draw on this to discuss how to interpret the simultaneous moderation of family income.

INTSPEC **Setup and** GFI **Analysis**

The *intspec* command is shown in boldface at the top of the Stata output for the GFI analysis below. I first define family income—*c.faminc*—as the focal variable in the focal( ) option. The main( ) option now specifies the three variables constituting the interaction, family income and its moderators birth cohort and education. Because *i.cohort* is a nominal factor variable, ICALC automatically includes all of its categories as display/calculation values, so the range( ) suboption would be ignored if it were listed. When you have two moderators, you should use the one with the fewest display categories as the second moderator for the GIF and SIGREG tools because it will define the columns of 2D tables of results. Given the limited column width in the Stata Results window, you will get easier-to-read displays this way. Note the ordering of the two-way interaction variables in the int2( ) option—*c.faminc#c.educ* is first and then *c.faminc#i.cohort*. Remember that this order must correspond to the relative order in which the moderators are listed in the main( ) option. (The focal variable can be listed anywhere in this ordering.)

In contrast to the prior examples, this is a more complicated interaction specification, with two moderators of family income, one of which is a multicategory nominal variable. So I use the *gfi* command to also produce a path–style diagram of the interaction effects in order to provide a visual display of the algebraic expressions. The suboptions for path( ) request the following:

- Show paths and coefficients for all interaction components in the graph (keyword "all").

- Format the coefficients in the diagram with four decimal places, like the other results.
- Title the diagram as "Interaction of …".

```
. intspec  focal(c.faminc10k)  ///
>        main((c.faminc10k , name(Family_Inc) range(1(3)19)) ///
>                (i.cohort , name(Cohort)) ///
>                (c.educ , name(Education) range(0(5)20)) ) ///
>        int2(c.faminc10k#i.cohort c.faminc10k#c.educ ) ndig(0)
   …

. gfi ,  ndig(4) path(all, ndig(4) boxw(1.5) ///
>     title("Interaction of Family Income by Cohort, Family Income by Education"))
```

```
GFI Information from Interaction Specification of
Effect of Family_Inc on g(Childs) from Linear Regression
----------------------------------------------------------------------

Effect of Faminc =
   -0.0487 - 0.0257*BabyBoom - 0.0282*WWII - 0.0645*DeprEra+ 0.0065*Educ

Sign Change Analysis of Effect of Family_Inc on g(childs)
 Moderated by Cohort (M1) and Education (M2)


--------------------------------------------------------------------------------
                          Effect of faminc10k
--------------------------------------------------------------------------------
             |       When Education =                              |
 When        | ---------------------------------------------------  |Sign Changes
 Cohort=     |  0          5          10         15         20      |   given M1
 ------------+-----------------------------------------------------+-----------
             |                                                      |
  PostBoom   |  Neg b=     Neg b=     Pos b=     Pos b=     Pos b=  | when M2=
             | -0.0487    -0.0161     0.0164     0.0490     0.0815  |    7.477
             |                                                      |
  BabyBoom   |  Neg b=     Neg b=     Neg b=     Pos b=     Pos b=  | when M2=
             | -0.0744    -0.0418    -0.0093     0.0232     0.0558  |   11.429
             |                                                      |
  WWII       |  Neg b=     Neg b=     Neg b=     Pos b=     Pos b=  | when M2=
             | -0.0769    -0.0443    -0.0118     0.0207     0.0533  |   11.814
             |                                                      |
  DeprEra    |  Neg b=     Neg b=     Neg b=     Neg b=     Pos b=  | when M2=
             | -0.1132    -0.0807    -0.0481    -0.0156     0.0169  |   17.396
             |                                                      |
-------------+-----------------------------------------------------+-----------
Sign Changes |    Never      Never  Sometimes  Sometimes      Never |
  given M2   |                                                      |
--------------------------------------------------------------------------------

Percent of in-sample cases with positive moderated effect of faminc10k = 74.0
```

Let's start with the GFI's algebraic expression of the moderated effect of family income; remember that income is measured in units of $10K. Because 0 is a valid value for the moderating variables, the main effect coefficient for family income

(−0.0487) has a meaningful interpretation. For someone in the Post-Boom cohort (reference category) who had no formal schooling (*Education* = 0), the number of children is predicted to decline by 0.05 children with a $10K increase in family income. When you have nominal moderators, it is useful to write out the algebraic expression separately for each category. You do this by substituting 0s and 1s into the cohort indicator variables[9] in the GFI expression to calculate the family income effect in each cohort. The Post-Boom cohort is 0 on all three indicators, and the remaining cohorts are coded 1 on their indicator and 0 on the other two. Applying this yields the following:

Post-Boom	$-0.0487-0.0257\times 0-0.0282\times 0-0.0645\times 0+0.0065\times \text{Educ}$
	$=-0.0487+0.0065\times \text{Educ}$
Baby Boom	$-0.0487-0.0257\times 1-0.0282\times 0-0.0645\times 0+0.0065\times \text{Educ}$
	$=-0.0744+0.0065\times \text{Educ}$
WWII	$-0.0487-0.0257\times 0-0.0282\times 1-0.0645\times 0+0.0065\times \text{Educ}$
	$=-0.0769+0.0065\times \text{Educ}$
DeprEra	$-0.0487-0.0257\times 0-0.0282\times 0-0.0645\times 1+0.0065\times \text{Educ}$
	$=-0.1132+0.0065\times \text{Educ}$

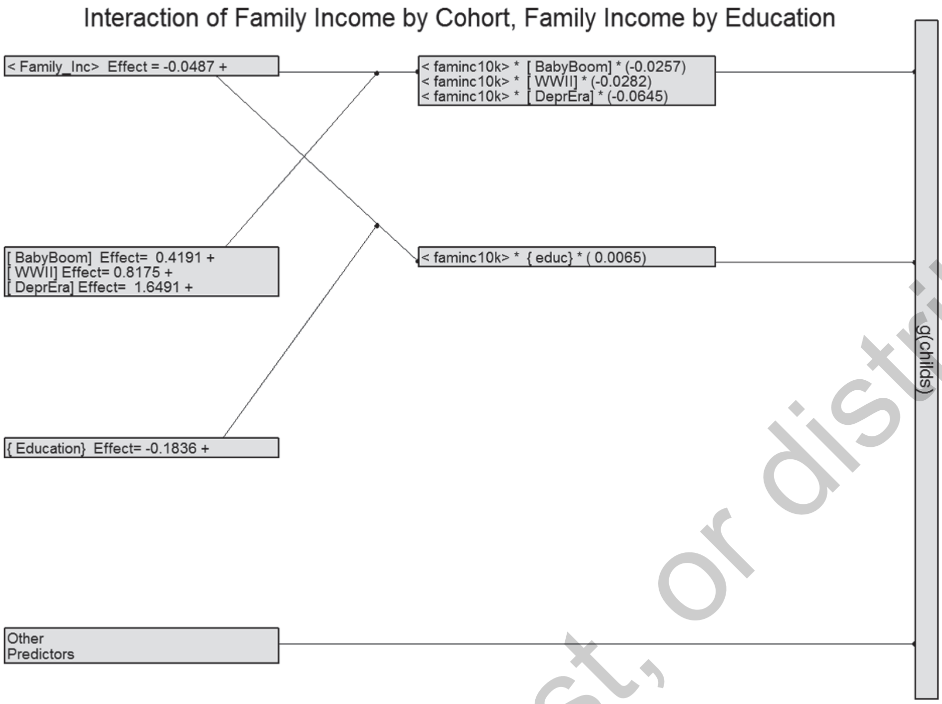
This calculation makes explicit the meaning of the GFI expression. The baseline negative effect of family income has its smallest magnitude for the Post-Boom cohort but grows larger for the successively older three birth cohorts (negative cohort coefficients), while the family income effect becomes less negative (more positive) as education increases (positive education coefficient). The path diagram in Figure 7.8 shows the structure of the interaction in terms of intersecting arrows leading from the left-hand column boxes to the second-column boxes. Family income and birth cohort interact because their arrows intersect, as do family income and education, but cohort and education do not interact (their arrows do not intersect), nor is there a three-way interaction of income, cohort, and education. The "Special Topics" section shows how to use the path diagram to read off the expressions above for the effect of income in each birth cohort.

The GFI expression raises the question of whether the family income effect changes from negative to positive as education increases and, if so, how that turnover point varies by birth cohort. The sign change analysis, the next part of the output, answers these questions. We can immediately see that the family income effect does in fact become positive at different values of education in each of the birth cohorts. The last column ("Sign Changes Given M1") indicates the sign change point for the birth cohorts. In the Post-Boom cohort, the family income effect becomes positive when *Education* > 7 years. For the Baby Boom and WWII cohorts, this change occurs when *Education* > 11 years, while for the Depression Era cohort, family income has a positive effect only for *Education* ≥ 18 years. The table note indicates that about three quarters of the estimation sample have a positive effect of family income on their number of children.

Significance Region Analyses: SIGREG and EFFDISP Tools

Especially because the family income effect changes sign, the next step is to explore its significance region to determine whether or not both positive and negative effects

FIGURE 7.8 ⬢ PATH-STYLE DIAGRAM OF INTERACTION EFFECTS MODEL



are significant. I do this with the *sigreg* command (at the top of the output) with no options because I want the default significance level (.05) and number of decimal places (four) to report the family income effect. A boundary values analysis is only possible for an interval moderator contingent on the other moderator(s). Thus, the output generates a warning message about boundary values for birth cohort contingent on education but reports the boundary values for education dependent on cohort. This analysis shows that for the Post-Boom and Baby Boom cohorts, the effect of family income is not significant unless education is greater than 13 and 14 years, respectively. In contrast, it is only significant for the Depression Era cohort for those with less than 8 years of education and is never significant for the WWII cohort.

Family Inc Effect Significance, Boundary Values for Education on **g(childs)** Given Cohort
Critical value F = **3.850** set with p = **0.0500**

Effect of Family_Inc	When Education >=	Sig Changes	When Education >=	Sig Changes
At Cohort =				
PostBoom	-43.4555 (< min)	to Not Sig [-0.0231]	13.4457	to Sig [1.6820]
BabyBoom	-13.9665 (< min)	to Not Sig [-0.0471]	14.8670	to Sig [2.5688]
WWII	-21.5109 (< min)	to Not Sig [-0.0396]	20.2875 (> max)	to Sig [0.6130]
DeprEra	7.3759	to Not Sig [-0.3525]	40.9869 (> max)	to Sig [0.0636]

Note: Derivatives of Boundary Values in [ ]

The boundary value results are primarily useful for providing detail when discussing the empirically derived significance region table and other results. This table shows the nature of the changing effect of family income more fully as it reports the sign, magnitude, and significance of the family income effect for each cohort at selected values of education. The formatting of the table makes it easy to see the two regions of significance:

- Positive and significant family income effects for the Post-Boom and Baby Boom cohorts with some college or higher education (more than 13 and 14 years, respectively)
- Negative and significant family income effects for the Depression Era cohort with less than 8 years of education

The significance region table and the content of these two bullet points provide a succinct and easy-to-understand portrayal of how family income predicts number of children differently for birth cohorts and for education levels. If I were to present this table, I would rerun *sigreg* with the `save( )` option to save it as the Excel formatted version shown in Table 7.4. With two dimensions in the table and the change in sign, the addition of the fill pattern highlighting is much more effective than the font highlighting in the output table.

Significance Region for Effect of faminc10k (1 unit difference)
on g(childs) at Selected Values of Cohort and Education

	At Cohort=			
At Education=	PostBoom	BabyBoom	WWII	DeprEra
0	_-0.0487_	_-0.0744_	_-0.0769_	_-0.1132*_
2	_-0.0356_	_-0.0614_	_-0.0639_	_-0.1002*_
4	_-0.0226_	_-0.0483_	_-0.0508_	_-0.0872*_
6	_-0.0096_	_-0.0353_	_-0.0378_	_-0.0742*_
8	0.0034	_-0.0223_	_-0.0248_	_-0.0611_
10	0.0164	_-0.0093_	_-0.0118_	_-0.0481_
12	0.0294	0.0037	0.0012	_-0.0351_
14	**0.0424***	0.0167	0.0142	_-0.0221_
16	**0.0555***	**0.0297***	0.0272	-0.0091
18	**0.0685***	**0.0428***	0.0403	0.0039
20	**0.0815***	**0.0558***	0.0533	0.0169

Key: Plain font = Pos, Not Sig   **Bold font*** = Pos, Sig
Italic font = Neg, Not Sig   _Italic font*_ = Neg, Sig

Alternatively, we can use the EFFDISP tool to create a set of plots of the family income effect on the *y*-axis against one of the moderators on the *x*-axis, repeated for each display value of the second moderator. A confidence bounds plot requires an interval moderator to define the *x*-axis, while an error bar plot typically uses a categorical moderator to define the *x*-axis. With a mixture of interval and categorical moderators, you get the information from a confidence bounds plot most efficiently. In this instance, we can see how the family income effect varies across all the values of education within a birth cohort, as well as how the income effect varies across all the birth cohort categories. An error bar plot would display the family income effect against all the birth cohorts, with plots for selected values of education; separate plots for the 21 distinct values of education would be a visual overload.

The EFFDISP tool always uses the first moderator to define the *x*-axis, but the *intspec* command used previously lists education as the second moderator. So I respecify *intspec* to list education first. In the *effdisp* command, the `plot( )` option generates

TABLE 7.4 ● EXCEL-FORMATTED SIGNIFICANCE REGION TABLE

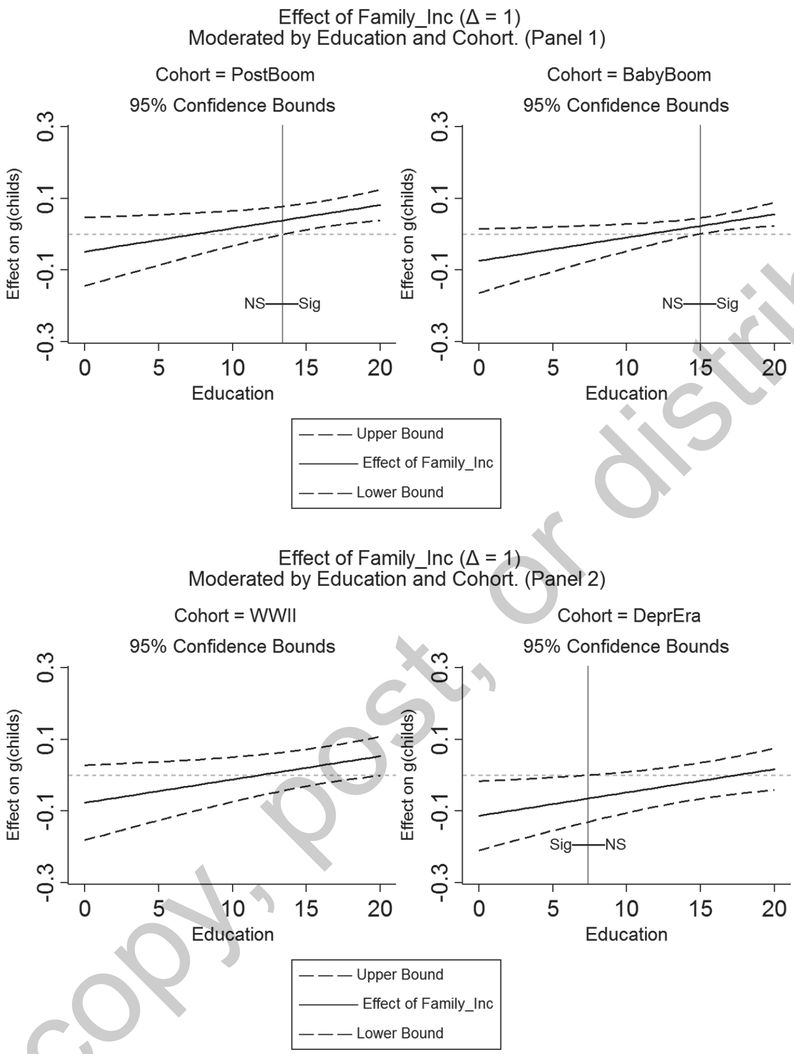
Effect of faminc10k (One-Unit Difference) Moderated by Cohort and Education on g(childs), Formatted to Highlight Sign and Significance				
	Cohort			
Education	**Post-Boom**	**Baby Boom**	**WWII**	**DeprEra**
0	*–0.0487*	*–0.0744*	*–0.0769*	***–0.1132****
2	*–0.0356*	*–0.0614*	*–0.0639*	***–0.1002****
4	*–0.0226*	*–0.0483*	*–0.0508*	***–0.0872****
6	*–0.0096*	*–0.0353*	*–0.0378*	***–0.0742****
8	0.0034	*–0.0223*	*–0.0248*	*–0.0611*
10	0.0164	*–0.0093*	*–0.0118*	*–0.0481*
12	0.0294	0.0037	0.0012	*–0.0351*
14	**0.0424***	0.0167	0.0142	*–0.0221*
16	**0.0555***	**0.0297***	0.0272	*–0.0091*
18	**0.0685***	**0.0428***	0.0403	0.0039
20	**0.0815***	**0.0558***	0.0533	0.0169
Key				
Plain font, no fill	Pos, Not Sig			
Bold*, filled	Pos, Sig			
Italic, no fill	Neg, Not Sig			
Bold italic*, filled	Neg, Sig			

confidence bounds plots with the specified name for saving them as memory graphs, and pltopts( ) formats the *y*-axis labels:

```
intspec focal(c.faminc10k) ///
      main((c.faminc10k, name(Family_Inc) range(1(3)19)) ///
            (c.educ, name(Education) range(0(5)20)) ///
            (i.cohort, name(Cohort))) ///
      int2( c.faminc10k#c.educ c.faminc10k#i.cohort) ndig(0)

effdisp, plot(type(cbound) name(FamInc)) ndig(1) pltopts(ylab(-.3(.2).3))
```

Figure 7.9 reports for each birth cohort the confidence bounds plots of the effect of family income by education. In each plot, the *solid line* is the moderated effect of family income, and the *dashed lines* show the confidence boundaries for the effect. The *thin gray horizontal reference line* separates negative and positive effects. This shows that the family income effect changes from negative to positive for each birth

FIGURE 7.9 ⬢ EFFDISP CONFIDENCE BOUNDS PLOT FOR TWO MODERATORS



cohort but at different values of education. A vertical reference line if present marks a change in the statistical significance of the family income effect.

For the Post-Boom and Baby Boom cohorts, it is apparent that the effect of family income is significant only at the upper end of the education distribution, where it has a positive effect. The effect of income is never significant for the WWII cohort, and for the Depression Era cohort, the income effect is negative and significant only when *Education* < 8 years. Like the significance region table, the confidence bounds plots provide an accessible and easy-to-interpret portrayal of the moderated effects of family income. In practice, I would also get error bar plots for family income by cohort to see how well they show the patterns of the income effect. In this instance, I think you would find that they also tell the story but it takes more effort to see the patterns.

Outcome Displays: OUTDISP Tool

A display of the predicted values can bring together all the components of the interaction specification. And the prior exploration of changes in the sign, magnitude, and statistical significance of the moderating effects provides a foundation for you to more easily see and interpret the patterns.

The ICALC commands are shown at the top of the output in boldface. I first respecify family income's display values on the *intspec* command (range(1(4.5)19)), with a larger increment between display values (every 4.5 units instead of every 3) to create a more compact predicted values table. The *outdisp* command has five options that create both a table and a scatterplot of the predicted number of children as it varies with family income, birth cohort, and education:

- out(atopt((means) _all)) sets the other predictors' reference values to their means.
- plot(name(Childs_by_FamInc_Cohort_Ed)) names the memory graph and produces the default plot, a scatterplot because the focal variable, family income, is interval.
- ndig(2) sets the format in the predicted values table to report two digits after the decimal.
- tab(def) creates the default table in which the focal variable defines the columns.
- pltopts( ) and its suboptions clean up the look of the initial default scatterplot. ylab and ymtick label the *y*-axis with values and ticks at 0, 2, 4, 6, and with tick marks between these values, respectively. xlab labels the *x*-axis reporting values with $1K scaling. ytit and xtit provide new titles for the *y*-axis and *x*-axis, respectively. See the "Special Topics" section for more details.

(*Note:* For an initial run, you could request results using the defaults very simply with outdisp plot(def) tab(def).)

```
. intspec  focal(c.faminc10k<)  ///
>        main((c.faminc10k , name(Family_Inc) range(1(4.5)19)) ///
>                (c.educ , name(Education) range(0(5)20))  ///
>                (i.cohort , name(Cohort))) ///
>        int2( c.faminc10k#c.educ c.faminc10k#i.cohort) ndig(1)
   …
.
. outdisp, out(atopt((means) _all)) plot(name(Childs_by_IncCohEd)) ndig(2) ///
>   tab(def) pltopts(ylab(0(2)6) ymtick(1 3 5) ytit("Number of Children") ///
>      xtit("Family Income") xlab(1 "$10K" 7 "$70K" 13 "$130K" 19 "$190K"))
…
```

Predicted Value of childs by the Two-way Interactions of Family_Inc with Education and with Cohort.

Cohort and Education	Family_Inc 1.0	5.5	10.0	14.5	19.0
PostBoom					
0	3.98	3.76	3.54	3.32	3.10
5	3.09	3.02	2.95	2.88	2.80
10	2.21	2.28	2.36	2.43	2.50
15	1.32	1.54	1.76	1.98	2.20
20	0.44	0.80	1.17	1.54	1.90
BabyBoom					
0	4.37	4.04	3.70	3.37	3.03
5	3.49	3.30	3.11	2.92	2.73
10	2.60	2.56	2.52	2.48	2.43
15	1.72	1.82	1.92	2.03	2.13
20	0.83	1.08	1.33	1.58	1.83
WWII					
0	4.77	4.42	4.08	3.73	3.38
5	3.88	3.68	3.48	3.28	3.08
10	3.00	2.94	2.89	2.84	2.78
15	2.11	2.20	2.30	2.39	2.48
20	1.23	1.47	1.70	1.94	2.18
DeprEra					
0	5.56	5.05	4.54	4.03	3.53
5	4.68	4.31	3.95	3.59	3.23
10	3.79	3.58	3.36	3.14	2.93
15	2.91	2.84	2.77	2.70	2.63
20	2.02	2.10	2.17	2.25	2.33

While the predicted values table is set up to facilitate the interpretation of the moderated effect of family income by education and birth cohort, it also can be used for interpreting how family income moderates education and how family income moderates cohort. To interpret the effect of education on number of children as moderated by income, we can use the panel of results for any cohort because they will show the identical pattern of magnitude differences in the effect of education. *This would not be true if there were a three-way interaction of income, education, and cohort, or if these results were for a nonlinear link model.* I use the Post-Boom cohort panel for convenience because it is the top one.

Reading down a column shows the effect of education on the predicted number of children for the given level of family income. Each column shows that the predicted number of children declines with education; for instance, at a family income of $55K (5.5), the predicted number of children drops from 3.76 for *Education* = 0 years to 0.80 for *Education* = 20, a difference of 2.96 children. Comparing the columns from left to right, the rate of decline with education decelerates as family income increases: −3.54, −2.96, −2.37, −1.78, and −1.20. And the effect of education on number of children is no longer significant once family income is greater than $170,529, as determined by the boundary value analysis. I would advise readers to verify for themselves that we get the same rate of decline, within rounding error, for the education effect from the other three panels.

Analogously, to interpret the effect of birth cohort, we can use any education row within a panel and compare it across the other panels. Let's use the bottom row in each panel for 20 years of education and start by comparing the Post-Boom cohort with each of the other cohorts. The predicted number of children for the Baby Boom cohort is larger than for the Post-Boom cohort at incomes between \$10K and \$145K (by 0.39, 0.28, 0.16, and 0.04) but is smaller—though not significant—than for the Post-Boom cohort at higher incomes (−0.07). For the other two (older) cohorts, the predicted number of children is always greater than for the Post-Boom cohort, but the differences diminish at higher levels of income and become nonsignificant (at \$115,750 for the WWII cohort and at \$152,130 for the Depression Era cohort). The overall pattern is a higher predicted number of children for older cohorts compared with younger cohorts, with the set of differences among the cohorts declining with family income and becoming nonsignificant at incomes greater than \$160K. The "Special Topics" section at the end of this chapter shows how to estimate the value of the moderator at which the differences in the predicted values among the categories of a nominal focal variable change significance.

Unlike the effect of birth cohort or education, the interpretation of family income's effect changes depending on which panel (cohort) and which row within a panel (level of education) you examine. That is, family income's effect is moderated by both cohort and education. Given the table organization, it is straightforward to discuss how family income's effect varies by education for each cohort. To talk more easily about specifics, I added a column to the predicted values table saved in Excel. The rightmost column in Table 7.5 shows the change in the number of children across the displayed range of income.

Within each cohort, the effect of family income is negative for low levels of education—the predicted number of children declines from left to right—and turns positive at higher levels of education—the predicted number of children rises from left to right. And across the cohorts, the younger the cohort, the smaller the magnitude of the negative effect of income and the larger the size of the positive effect. In the Post-Boom cohort, family income's effect changes to a positive effect when *Education* > 7.48 but does not become significant until *Education* > 13.45. Its initial negative slope is the shallowest, with a change of −0.88 in the number of children, and its final positive slope is the steepest among the cohorts at almost 1.5 children. The changeover points are higher for the effect of family income in the Baby Boom than in the Post-Boom cohort, 11.43 for the sign change and 14.87 for the significance change. And correspondingly, its most negative slope is steeper (–1.34) and its most positive slope shallower (1.00) than for the Post-Boom cohort. The WWII cohort has a very similar sign changeover point in slopes as the Baby Boom cohort (11.81), but the effect of family income does not become significant for any level of education. For the Depression Era cohort, the family income effect stays negative for all but the highest levels of education (>17.40). The negative effect is significant when *Education* < 7.4, but its positive effect is never significant. Not surprisingly, it has the largest magnitude negative slope (–2.04) and the smallest positive slope (0.31).

A good alternative is to use a predicted values plot, for which you do not necessarily need to do side calculations because the magnitudes of the slope for family income are visually apparent. But it can be useful to do so as a check that your visual assessment of steeper and shallower slopes is accurate. Figure 7.10 presents a scatterplot for each cohort, in which the predicted number of children is plotted against family income with separate lines for selected values of education. The *solid line*, which is the highest in each plot, represents the effect of family income for *Education* = 0, and the successive lower lines represent the income effect for progressively higher education levels.

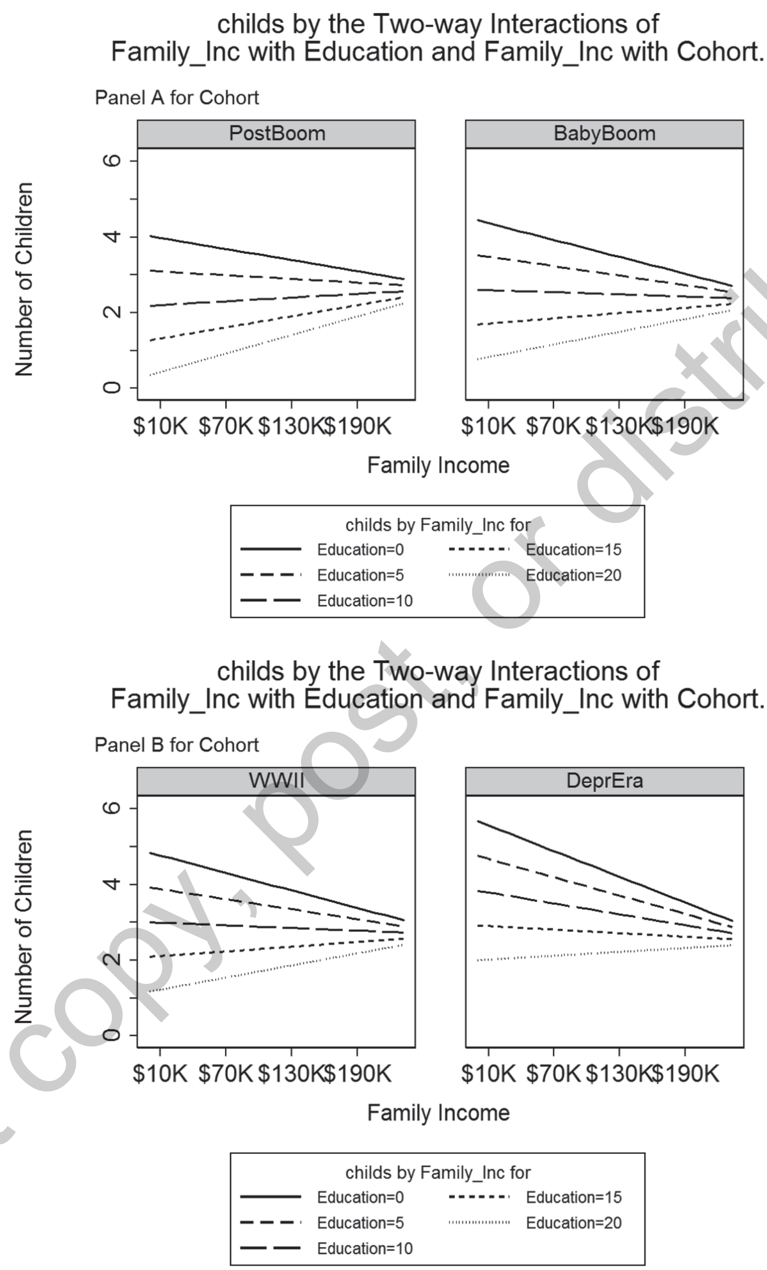
TABLE 7.5 ⬢ OUTDISP PREDICTED VALUES TABLE WITH ADDED CALCULATION

childs by the Two-Way Interactions of Family_Inc With Education and With Cohort							
Cohort	**Education**	**Family Income**					**Change in Number of Children**
Post-Boom	0	3.98	3.76	3.54	3.32	3.10	−0.88
	5	3.09	3.02	2.95	2.88	2.80	−0.29
	10	2.21	2.28	2.36	2.43	2.50	0.30
	15	1.32	1.54	1.76	1.98	2.20	0.88
	20	0.44	0.80	1.17	1.54	1.90	1.47
Baby Boom	0	4.37	4.04	3.70	3.37	3.03	−1.34
	5	3.49	3.30	3.11	2.92	2.73	−0.75
	10	2.60	2.56	2.52	2.48	2.43	−0.17
	15	1.72	1.82	1.92	2.03	2.13	0.42
	20	0.83	1.08	1.33	1.58	1.83	1.00
WWII	0	4.77	4.42	4.08	3.73	3.38	−1.38
	5	3.88	3.68	3.48	3.28	3.08	−0.80
	10	3.00	2.94	2.89	2.84	2.78	−0.21
	15	2.11	2.20	2.30	2.39	2.48	0.37
	20	1.23	1.47	1.70	1.94	2.18	0.96
DeprEra	0	5.56	5.05	4.54	4.03	3.53	−2.04
	5	4.68	4.31	3.95	3.59	3.23	−1.45
	10	3.79	3.58	3.36	3.14	2.93	−0.87
	15	2.91	2.84	2.77	2.70	2.63	−0.28
	20	2.02	2.10	2.17	2.25	2.33	0.31

Note: WWII = World War II; DeprEra = Depression Era.

These plots clearly show the pattern just described above. The family income effect changes from negative to positive as education increases—compare the lines at the top of each plot with those at the bottom—and the steepness of the slopes differs by birth cohort. The Post-Boom cohort has the steepest positive slopes and shallowest negative slopes, and the Depression Era cohort has the shallowest positive slopes and steepest negative slopes, with the other two cohorts similarly in between. Also notice that the angles between the lines in each plot are identical for each cohort, reflecting the fact that the model does not include a three-way interaction of income, education, and cohort.

FIGURE 7.10 ⬢ OUTDISP SCATTERPLOT FOR TWO MODERATORS



The potential drawback of using scatterplots is that they do not lend themselves to describing the effects of education quite as easily or, especially, the effects of cohort. To interpret the education effect (which has the same pattern for each cohort), we can look at the vertical placement of each line as well as how the vertical distance between the plotted lines changes with family income. Because the plotted lines are vertically ordered, with the lines for smaller versus larger values of education having a higher predicted number of children at any value of family income, we know that the effect of education is negative. Furthermore, the gap between the lines diminishes at higher levels of family income, which indicates that the negative effect of education decreases in magnitude as family income increases. From the boundary value analysis, we know that the education effect becomes not significant for incomes greater than $152K. While this provides an accurate description of the education effect, it is harder to explain (and for readers to understand) than using a table or plot designed to highlight the effect of education.

To interpret the effect of birth cohort, use the line for any level of education, and compare it across the four plots. Looking at the left end of the top line for 0 years of education, we can see that the predicted number of children is highest in the Depression Era cohort and smallest in the Post-Boom cohort. But it is not possible to say with any confidence whether the Baby Boom cohort or the WWII cohort has the highest predicted number of children in the scatterplots. We can also deduce that the differences among the cohorts are diminishing. The lines with the steeper decreases in predicted values correspond to the cohorts with the higher initial predicted number of children. Moreover, it appears that the right-hand end of the *solid lines* for the cohorts is less spread out in vertical location than the left-hand ends of the lines. All in all, this is not a very compelling or informative way to describe the moderated effect of birth cohort.

What to Present and Interpret?

Let me start by reemphasizing that you would only present a very limited subset of the results from the analyses we applied to this example. Most are intended to help you, the analyst, better understand the interaction effects so that you can then better explain it to your audience. I outline below what I think are the best options for this example, but these are not necessarily what might work best in other analyses or what might convey the information best for you. Keep in mind that you should discuss how each of the three component variables of the interaction—family income, education, and birth cohort—are related to the outcome. But what works well for interpreting one of these may not work as well for another. For example, the scatterplot in Figure 7.10 is very useful for interpreting the family income effect, while it is difficult to use for the effect of birth cohort. And, as you will see in the next chapter, what works well for a linear link model may not work equally well for a nonlinear link model.

I think that the table of predicted values with the added column of the predicted change in the number of children across the range of income (Table 7.5) is very accessible since it reports predicted values in an intuitive metric: the number of children. And it provides an effective portrayal of how the predicted number of children varies with all three component variables of the interaction. It is effective because the patterns are not complicated and this is a linear link model. It is easy to supplement it with information from the GFI and significance region analyses and can be concisely described. The relationship between number of children and education or between number of children and birth cohort can be described in a short paragraph each. The discussion of the family income effect is longer because it describes how the effect is contingent on two other factors.[10] It also has the advantage that you can use a single display (table) for all three components of the interaction. For added visual impact, you could present the font-size-proportional-to-value-size version of the table, as in Table 7.3.

A second, very good option would be to use three graphic displays of the predicted values, one for each of the variables in the interaction specification, because each graphic would highlight the effects of a different variable on the outcome. That is, I would use scatterplots for family income (Figure 7.10) and for education (Figure 7.7) but a bar chart for birth cohort (Figure 7.6). Although the figures require more space, they are a visually appealing, intuitive, and effective presentation of the relationships for linear link models. And they similarly lend themselves to succinct discussions.

A third good option would be a formatted significance region table for each of the component variables. These highlight changes in the sign and significance of the effects of the component variables and also show the varying magnitude of the effect. For education and family income, I would recommend calculating and reporting the effects for a standard deviation change in the predictor rather than a one-unit change. This would put both predictors on a common scale—standard deviation units—and let you compare the relative magnitude of their effects on the number of children. Depending on your audience, this may take more explanation and be less accessible than relying on the predicted values display, whether tabular or graphical.

Combinations of these three types can also be a quite reasonable option—for example, presenting the predicted values table for the family income effect as well as a bar chart showing the relationship between number of children and birth cohort (Figure 7.6), or presenting formatted significance region tables for the effect of birth cohort and for the effect of education but presenting a predicted values scatterplot for the effect of family income. You may have noticed that I did not include effects displays (confidence bounds plots or error bar plots) among my recommendations. Although I find these visually appealing, I think that many readers are initially confused by them as they expect to see plots of predicted values.

Looking ahead, this chapter concludes with material on several specialized topics or details that would have unduly interrupted the flow of discussion. The next chapter begins our tour through some of the most common GLMs with nonlinear link functions and how to address the challenges of interpretation in those contexts. We begin with binomial logistic regression and probit analysis in Chapter 8.

SPECIAL TOPICS

Customizing Plots With the pltopts( ) Option

Consider the scatterplot in the upper panel of Figure 7.7. Although it is visually easy to interpret, the axis labeling is not ideal. You can create a much cleaner appearance either by using the graph editor in Stata to make changes or by rerunning the *outdisp* command and adding the pltopts( ) option to modify the appearance of the scatterplot. The lower panel was produced with the following command:

```
outdisp, out(atopt( (means) _all)) plot(name(Plot2)) pltopts( xlab(0(4)20) ///
      ylab(1 3 5) ymtick(2 4, tl(**2) ysc(r(.5)) )
```

The content of pltopts( ) consists of Stata two-way graph options that do the following:

- *xlab(0(4)20)* replaces the education value labels on the *x*-axis with their numeric values.

- *ylab(1 3 5)* replaces the noninteger labels for the number of children on the *y*-axis with whole number labels for 1, 3, and 5 children.
- *ymtick(2 4, tl(*2))* adds minor tick marks on the *y*-axis between the new labels at values of 2 and 4, and tl(*2) makes the tick marks twice their usual length.
- *ysc(r(.5))* extends the *y*-axis to a value of 0.5 without a tick mark or a label, to roughly equalize the top- and bottom-margin areas above and below the plotted lines, respectively.

Additional examples of how you can customize plots are shown in the *outdisp* command used to create Figure 7.10, which changes the titles for the *y*- and *x*-axes and adds custom value labels to the *x*-axis. It is beyond the scope of this book to review and explain all of the two-way options and how they might be used to modify plots created with ICALC, including when they won't work. The options are described in the Stata documentation, but I would recommend Mitchell's (2012) book on Stata graphics as an excellent reference on the graph commands and their options, as well as on using the graph editor to customize an already drawn plot. Alternatively, you could specify the save( ) suboption within plot( ) in the *outdisp* command. This will save the data used to create the plot in an Excel file, and then you can construct and customize your own graphics with whatever software platform you prefer.

Aside on Using the Path Diagram for a Multicategory Nominal Moderator

When you have nominal moderators, it is useful to write out the algebraic expression separately for each category. In the discussion of the GFI results for family income moderated by birth cohort and education, I showed how to do this algebraically. You can also read these directly from the path diagram of the interaction model, as I described in Chapter 6, repeating the process for each birth cohort. Start with the family income effect in the top box of the left-hand column, and trace each arrow to the boxes in the second column. That is, follow the horizontal arrow to the top box of the second column, and add the relevant coefficient for the birth cohort, dropping both variable labels. Then go back and follow the diagonal arrow to the other box in the second column, and add its contents, dropping the variable label for family income. For instance, for the WWII birth cohort the family income effect is equal to

−0.0487	from the top left-hand box
−0.0282	from horizontal arrow to 2nd column, top box, Line 2
+0.0065 × *Educ*	from diagonal arrow to 2nd column, 2nd box

giving −0.0769 + 0.0065 × *Educ*. For the base birth cohort (Post-Boom), there is no relevant coefficient in the second column's top box, so you add nothing to its expression and proceed to tracing the next arrow.

Testing Differences in the Predicted Outcome Among Categories of a Nominal Variable

I used an iterative computational approach to determine the value of family income at which the differences in the predicted number of children among the birth cohorts were not significantly different from 0. I relied on the *mtable* and *mlincom* commands in SPOST13 for doing the calculations. After describing the step-by-step process, I list a simple program that can be run to do the calculations; it is also available to download from www.icalcrlk.com.

The coded income values ranged from 0.5 to 19.2. So for the first iteration, I generated the predicted number of children for the birth cohorts for the 21 integer values of income from 0 to 20 with the *mtable* command:

```
mtable, at (faminc10k = (0(1)20) cohort=(1/4)) atmeans stat(pvalue noci) post
```

This produces 21 sets of estimates for each cohort, for a total of 84 stored estimates. For each set, I tested if the sum of the differences between the estimates for each unique pair of cohorts is equal to 0. With four cohorts, there are six unique pairs of cohorts to compare:

$$\left(\hat{y}_{cohort\,2} - \hat{y}_{cohort\,1}\right) + \left(\hat{y}_{cohort\,3} - \hat{y}_{cohort\,1}\right) + \left(\hat{y}_{cohort\,4} - \hat{y}_{cohort\,1}\right)$$
$$+\left(\hat{y}_{cohort\,3} - \hat{y}_{cohort\,2}\right) + \left(\hat{y}_{cohort\,4} - \hat{y}_{cohort\,2}\right) + \left(\hat{y}_{cohort\,4} - \hat{y}_{cohort\,3}\right) = 0$$

I performed the test with the *mlincom* command by referring to the position of the stored estimates. For example, the second set of estimates for the birth cohorts (when *faminc* = 1) is stored in positions 5 to 8, and the *mlincom* command is

```
mlincom 6-5 + 7-5 + 8-5 + 7-6 + 8-6 + 8-7
```

This produced a test statistic value of 5.149 with a p value of .000. I repeated this calculation for all 21 sets of estimates. This showed that the difference was significant at a family income value of 15 (p = .022) but not significant at the value of 16 (p = .051).

The second iteration increased the precision by one order of magnitude (to the nearest 0.1). I followed the same process to test the differences among cohorts for family income coded values from 15 to 16 in increments of 0.1. This showed that the intercohort differences were first not significant at a family income of 16.0, which represents a value of $160K (16.0 × $10,000). I could have done additional iterations with smaller increments to get additional digits for a more precise income value at which the differences turned nonsignificant.

In reality, I used the program shown in the box to automate each iteration, but it only applies to testing a nominal variable that is moderated by one other variable. After the program is loaded in your Stata session,[11] you access it with the following command:

mcattest mcvar(*nominal variable name*) var2(*moderator name*) vallist (*numlist*)

The numlist must be of the form #1/#2 or #1(#2)#3.

The command for the first iteration is

```
mcattest, mcvar(cohort) var2(faminc10k) vallist(0(1)20)
```

and for the second iteration,

```
mcattest, mcvar(cohort) var2(faminc10k) vallist(15(.1)16)
```

```
program mcattest
syntax, mcvar(varname) var2(varname) vallist(string)
tempname estnm
qui{
est store `estnm'
levelsof `mcvar', loc(nval)
loc ncat : list sizeof nval
mtable , at (`var2' "= (`vallist')" `mcvar'=(`nval')) atmeans stat(pvalue noci) post
mlincom, clear
loc atind = 1-`ncat'
forvalues fi = `vallist' {
    loc atind= `atind' + `ncat'
    loc difftxt ""
    forvalues i=1/`=`ncat'-1' {
    forvalues j=1/`=`ncat'- `i'' {
        if `j' == 1 & `i'== 1 loc difftxt "`difftxt' `=`atind'+`j'' - `atind' "
        if `j' > 1 | `i' > 1 loc difftxt "`difftxt' + `=`atind'+`j'+`i'-1' - `=`atind'+`i'-1' "
    }
    }
    mlincom `difftxt', add rowname("`fi'")
}
}
mlincom
qui est restore `estnm'
end
```


CHAPTER 7 NOTES

1. These models are commonly labeled as GLS because they generalize (relax) the assumptions of equal error variance and zero covariance. I avoid using this label in the text exposition in order to avoid the inevitable confusion about what is meant by labeling a model GLS versus GLM or by making statements such as "a GLS model is one type of GLM."
2. Response categories with a range of values used the midpoint value. The top-coded category for frequency of sex (4+ times per week) was recoded using a value of 4.5 times

per week. Redoing the analyses using a value of 5 times per week produced essentially identical results.

3. In all, 269 cases were missing on frequency of sex, 109 more were missing on attendance at religious services, and 53 more were missing on education.
4. The mean of SES in the sample analyzed is 53.2, which is a bit larger than the mean in the full sample of 52.2.
5. This is specified for didactic reasons. For OLS regression, the *margins* command produces exactly the same results for the "as observed" method of treating the predictors and for the "as means" option. Thus, in practice, this option is unnecessary for OLS analyses.
6. The birth cohorts are defined by respondents' age in 2010: Depression Era (70 years or older), World War II (65–69), Baby Boom (46–64), and Post-Boom (40–45). Translated to year born, the categories are Depression Era (before 1941), World War II (1941–1945), Baby Boom (1946–1964), and Post-Boom (1965–1970).
7. In all, 172 cases were missing on income, an additional 45 were missing on religious intensity, and 1 more was missing on education.
8. These cases all had high leverage values but moderate residuals. They all had the maximum (top coded) value for income, were in the two oldest birth cohorts, and had an atypically small number of children. Removing these cases from the analysis had negligible effects on the estimation results. Hence, the analyses include them.
9. If your nominal variable is contrast coded rather than dummy coded, you would instead substitute the contrast-code values to specify a given cohort.
10. Taking out my commentary from what I wrote above, the description for education was 176 words in five sentences, and for cohort, 117 words in four sentences.
11. mcattest.do can be downloaded from icalcrlk.com from the icalc_spec package. You could either highlight the set of lines for the program and click on the run icon or issue a run command specifying the path and the filename—run *path*/mcattest.do. It is also listed at the end of the downloadable do-file for this example.