Interpreting and Visualizing Regression Models Using Stata

Michael N. Mitchell

A Stata Press Publication
StataCorp LP
College Station, Texas

Copyright © 2012 by StataCorp LP
All rights reserved. First edition 2012

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in $\LaTeX\,2_\varepsilon$
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-107-2
ISBN-13: 978-1-59718-107-5

Library of Congress Control Number: 2011943975

No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without the prior written permission of StataCorp LP.

Stata, stata, Stata Press, Mata, mata, and NetCourse are registered trademarks of StataCorp LP.

Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations.

$\LaTeX\,2_\varepsilon$ is a trademark of the American Mathematical Society.

Contents

(Pages omitted)

Preface

Think back to the first time you learned about simple linear regression. You probably learned about the underlying theory of linear regression, the meaning of the regression coefficients, and how to create a graph of the regression line. The graph of the regression line provided a visual representation of the intercept and slope coefficients. Using such a graph, you could see that as the intercept increased, so did the overall height of the regression line, and as the slope increased, so did the tilt of the regression line. Within Stata, the `graph twoway lfit` command can be used to easily visualize the results of a simple linear regression.

Over time we learn about and use fancier and more abstract regression models—models that include covariates, polynomial terms, piecewise terms, categorical predictors, interactions, and nonlinear models such as logistic. Compared with a simple linear regression model, it can be challenging to visualize the results of such models. The utility of these fancier models diminishes if we have greater difficulty interpreting and visualizing the results.

With the introduction of the `marginsplot` command in Stata 12, visualizing the results of a regression model, even complex models, is a snap. As implied by the name, the `marginsplot` command works in tandem with the `margins` command by plotting (graphing) the results computed by the `margins` command. For example, after fitting a linear model, the `margins` command can be used to compute adjusted means as a function of one or more predictors. The `marginsplot` command graphs the adjusted means, allowing you to visually interpret the results.

The `margins` and `marginsplot` commands can be used following nearly all Stata estimation commands (including `regress`, `anova`, `logit`, `ologit`, and `mlogit`). Furthermore, these commands work with continuous linear predictors, categorical predictors, polynomial (power) terms, as well as interactions (for example, two-way interactions, three-way interactions). This book uses the `marginsplot` command not only as an interpretive tool, but also as an instructive tool to help you understand the results of regression models by visualizing them.

Categorical predictors pose special difficulties with respect to interpreting regression models, especially models that involve interactions of categorical predictors. Categorical predictors are traditionally coded using dummy (indicator) coding. Many research questions cannot be answered directly in terms of dummy variables. Furthermore, interactions involving dummy categorical variables can be confusing and even misleading. Stata 12 introduces the `contrast` command, a general-purpose command that can be used to precisely test the effects of categorical variables by forming contrasts among the levels of the categorical predictors. For example, you can compare adjacent groups, compare each group with the overall mean, or compare each group with the mean of the previous groups. The `contrast` command allows you to easily focus on the comparisons that are of interest to you.

The `contrast` command works with interactions as well. You can test the simple effect of one predictor at specific levels of another predictor or form interactions that involve comparisons of your choosing. In the parlance of analysis of variance, you can test simple effects, simple contrasts, partial interactions, and interaction contrasts. These kinds of tests allow you to precisely understand and dissect interactions with surgical precision. The `contrast` command works not only with the `regress` command, but also with commands such as `logit`, `ologit`, `mlogit`, as well as random-effects models like `xtmixed`.

As you can see, the scope of the application of the `margins`, `marginsplot`, and `contrast` commands is broad. Likewise, so is the scope of this book. It covers continuous variables (modeled linearly, using polynomials, and piecewise), interactions of continuous variables, categorical predictors, interactions of categorical predictors, as well as interactions of continuous and categorical predictors. The book also illustrates how the `margins`, `marginsplot`, and `contrast` commands can be used to interpret results from multilevel models, models where time is a continuous predictor, models with time as a categorical predictor, nonlinear models (such as logistic regression or ordinal logistic regression), and analyses that involve complex survey data. However, this book does not contain information about the theory of these statistical models, how to perform diagnostics for the models, the formulas for the models, and so forth. The summary section concluding each chapter includes references to books and articles that provide background for the techniques illustrated in the chapter.

My goal for this book is to provide simple and clear examples that illustrate how to interpret and visualize the results of regression models. To that end, I have selected examples that illustrate large effects generally combined with large sample sizes to create patterns of effects that are easy to visualize. Most of the examples are based on real data, but some are based on hypothetical data. In either case, I hope the examples help you understand the results of your regression models so you can interpret and present them with clarity and confidence.

Simi Valley, California Michael N. Mitchell
March 2012

(Pages omitted)

14 Continuous by categorical by categorical interactions

14.1 Chapter overview

This chapter considers models that involve the interaction of two categorical predictors with a linear continuous predictor. Such models blend ideas from chapter 10 on categorical by continuous interactions and ideas from chapter 8 on categorical by categorical interactions. As we saw in chapter 10, interactions of categorical and continuous predictors describe how the slope of the continuous variable differs as a function of the categorical variable. In chapter 8, we saw models that involve the interaction of two categorical variables. This chapter blends these two modeling techniques by exploring how the slope of the continuous variable varies as a function of the interaction of the two categorical variables.

Let's consider a hypothetical example of a model with income as the outcome variable. The predictors include gender (a two-level categorical variable), education (treated as a three-level categorical variable), and age (a continuous variable). Income can be modeled as a function of each of the predictors, as well as the interactions of all the predictors. A three-way interaction of age by gender by education would imply that the effect of age interacts with gender by education. One way to visualize such an interaction would be to graph age on the x axis, with separate lines for the levels of education and separate graphs for gender. Figure 14.1 shows such an example, illustrating how the slope of the relationship between income and age varies as a function of education and gender.

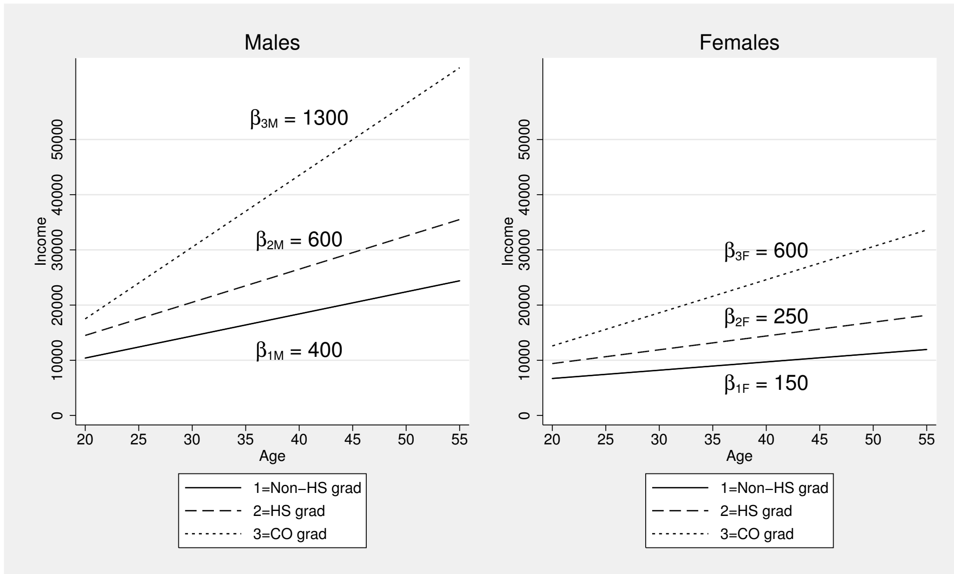

Figure 14.1. Fitted values of income as a function of age, education, and gender

The graph can be augmented by a table that shows the `age` slope broken down by education and gender. Such a table is shown in 14.1. The `age` slope shown in each cell of table 14.1 reflects the slope of the relationship between income and age for each of the lines illustrated in figure 14.1. For example, β_{3M} represents the `age` slope for male college graduates, and this slope is 1,300.

Table 14.1. The `age` slope by level of education and gender

	Non-HS grad	HS grad	CO grad
Male	$\beta_{1M} = 400$	$\beta_{2M} = 600$	$\beta_{3M} = 1{,}300$
Female	$\beta_{1F} = 150$	$\beta_{2F} = 250$	$\beta_{3F} = 600$

The age by education by gender interaction described in table 14.1 can be understood and dissected like the two by three interactions illustrated in chapter 8. The key difference is that table 14.1 is displaying the slope of the relationship between income and age, and the three-way interaction refers to the way that the slope varies as a function of education and gender.[1]

If there were no three-way interaction of age by gender by education, we would expect (for example) that the gender difference in the `age` slope would be approximately the

1. More precisely, how the slope varies as a function of the interaction of age and gender.

same at each level of education. But, consider the differences in the age slopes between females and males at each level of education. This difference is -250 $(150 - 400)$ for non–high school graduates, whereas this difference is -350 $(250 - 600)$ for high school graduates, and the difference is -700 $(600{-}1300)$ for college graduates. The difference in the age slopes between females and males seems to be much larger for college graduates than for high school graduates and non–high school graduates. This pattern of results appears consistent with a three-way interaction of age by education by gender.

Let's explore this in more detail with an example using the GSS dataset. To focus on the linear effect of age, we will keep those who are 22 to 55 years old.

```
. use gss_ivrm
. keep if age>=22 & age<=55
(18936 observations deleted)
```

In this example, let's predict income as a function of gender (female), a three-level version of education (educ3), and age. The regress command below predicts realrinc from i.female, i.educ3, and c.age (as well as all interactions of the predictors). The variable i.race is also included as a covariate.

```
. regress realrinc i.female##i.educ3##c.age i.race, vce(robust) vsquish
Linear regression                               Number of obs =   25718
                                                F( 13, 25704) =  411.30
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1839
                                                Root MSE      =   23556
```

realrinc	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.female	1337.125	1693.694	0.79	0.430	-1982.61	4656.861
educ3						
2	550.476	1782.192	0.31	0.757	-2942.721	4043.673
3	-11156.1	2618.976	-4.26	0.000	-16289.44	-6022.756
female#educ3						
1 2	783.0991	2021.654	0.39	0.698	-3179.457	4745.655
1 3	7657.907	3164.299	2.42	0.016	1455.703	13860.11
age	413.8695	45.62015	9.07	0.000	324.4515	503.2876
female#c.age						
1	-264.9842	50.65695	-5.23	0.000	-364.2746	-165.6937
educ3#c.age						
2	175.8497	54.7504	3.21	0.001	68.53584	283.1636
3	897.3326	77.47101	11.58	0.000	745.4851	1049.18
female# educ3#c.age						
1 2	-80.30545	60.94575	-1.32	0.188	-199.7625	39.15165
1 3	-414.6562	93.26714	-4.45	0.000	-597.465	-231.8473
race						
2	-2935.138	273.3294	-10.74	0.000	-3470.879	-2399.397
3	185.3956	956.338	0.19	0.846	-1689.081	2059.872
_cons	2691.23	1495.778	1.80	0.072	-240.5797	5623.039

Let's test the interaction of gender, education, and age using the `contrast` command below. The three-way interaction is significant.

```
. contrast i.female#i.educ3#c.age
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
female#educ3#c.age	2	10.17	0.0000
Residual	25704		

To begin the process of interpreting the three-way interaction, let's create a graph of the adjusted means as a function of age, education, and gender. First, the `margins` command below is used to compute the adjusted means by gender and education for ages 22 and 55 (the output is omitted to save space). Then the `marginsplot` command is used to graph the adjusted means, as shown in figure 14.2.

```
. margins female#educ3, at(age=(22 55))
  (output omitted)
. marginsplot, bydimension(female) noci
  Variables that uniquely identify margins: age female educ3
```

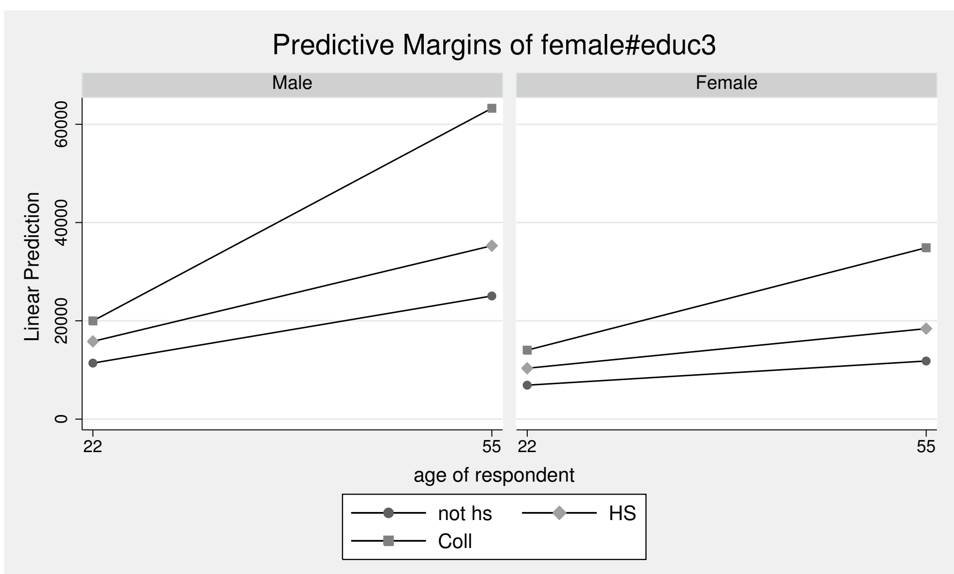

Figure 14.2. Fitted values of income as a function of age, education, and gender

The graph in figure 14.2 illustrates how the `age` slope varies as a function of gender and education. Let's compute the `age` slope for each of the lines shown in this graph. The `margins` command is used with the `dydx(age)` and `over()` options to compute the `age` slopes separately for each combination of gender and education.

```
. margins, dydx(age) over(female educ3)

Average marginal effects                        Number of obs   =      25718
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age
over         : female educ3

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
female#educ3 |
        0 1  |   413.8695   45.62015     9.07   0.000     324.4557    503.2834
        0 2  |   589.7192   30.37993    19.41   0.000     530.1757    649.2628
        0 3  |   1311.202   62.88374    20.85   0.000     1187.952    1434.452
        1 1  |   148.8854   22.09037     6.74   0.000      105.589    192.1817
        1 2  |   244.4296   15.25412    16.02   0.000     214.5321    274.3272
        1 3  |   631.5618   46.90854    13.46   0.000     539.6227    723.5008
------------------------------------------------------------------------------
```

Let's reformat the output of the `margins` command to emphasize how the `age` slope varies as a function of the interaction of gender and education (see table 14.2). Each cell of table 14.2 shows the `age` slope for the particular combination of gender and education. For example, the `age` slope for males with a college degree is 1,311.20 and is labeled as β_{3M}.

Table 14.2. The `age` slope by level of education and gender

	Non-HS grad	HS grad	CO grad
Male	$\beta_{1M} = 413.87$	$\beta_{2M} = 589.72$	$\beta_{3M} = 1{,}311.20$
Female	$\beta_{1F} = 148.89$	$\beta_{2F} = 244.43$	$\beta_{3F} = 631.56$

We can dissect the three-way interaction illustrated in table 14.2 using the techniques from section 8.3 on two by three models. Specifically, we can use simple effects analysis, simple contrasts, and partial interactions.

14.2 Simple effects of gender on the age slope

We can use the `contrast` command to test the simple effect of gender on the `age` slope. This is illustrated below.

```
. contrast female#c.age@educ3, nowald pveffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	Contrast	Std. Err.	t	P>|t|
female@educ3#c.age				
(1 vs base) 1	-264.9842	50.65695	-5.23	0.000
(1 vs base) 2	-345.2896	33.98931	-10.16	0.000
(1 vs base) 3	-679.6404	78.4498	-8.66	0.000

Each of these tests represents the comparison of females versus males in terms of the age slope. The first test compares the age slope for females versus males among non–high school graduates. Referring to table 14.2, this test compares β_{1F} with β_{1M}. The difference in these age slopes is -264.98 $(148.89 - 413.87)$, and this difference is significant. The age slope for females who did not graduate high school is 264.98 units smaller than the age slope for males who did not graduate high school. The second test is similar to the first, except the comparison is made among high school graduates, comparing β_{2F} with β_{2M} from table 14.2. This test is also significant. The third test compares the age slope between females and males among college graduates (that is, comparing β_{3F} with β_{3M}). This test is also significant. In summary, the comparison of the age slope for females versus males is significant at each level of education.

14.3 Simple effects of education on the age slope

We can also look at the simple effects of education on the age slope at each level of gender. This test is performed using the contrast command below.

```
. contrast educ3#c.age@female
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	df	F	P>F
educ3@female#c.age			
0	2	70.96	0.0000
1	2	43.37	0.0000
Joint	4	57.21	0.0000
Residual	25704		

The first test compares the age slope among the three levels of education for males. Referring to table 14.2, this tests the following null hypothesis.

$H_0\colon\ \beta_{1M} = \beta_{2M} = \beta_{3M}$

This test is significant. The age slope significantly differs as a function of education among males.

The second test is like the first test, except that the comparisons are made for females. This tests the following null hypothesis.

$H_0\colon\ \beta_{1F} = \beta_{2F} = \beta_{3F}$

This test is also significant. Among females, the `age` slope significantly differs among the three levels of education.

14.4 Simple contrasts on education for the age slope

We can further dissect the simple effects tested above by applying contrast coefficients to the education factor. For example, say that we used the `ar.` contrast operator to form reverse adjacent group comparisons. This would yield comparisons of group 2 versus 1 (high school graduates with non–high school graduates) and group 3 versus 2 (college graduates with high school graduates). Applying this contrast operator yields simple contrasts on education at each level of gender, as shown below.

```
. contrast ar.educ3#c.age@female, nowald pveffects
Contrasts of marginal linear predictions
Margins      : asbalanced
```

	Contrast	Std. Err.	t	P>|t|
educ3@female#c.age				
(2 vs 1) 0	175.8497	54.7504	3.21	0.001
(2 vs 1) 1	95.54426	26.83611	3.56	0.000
(3 vs 2) 0	721.4829	69.74939	10.34	0.000
(3 vs 2) 1	387.1322	49.38976	7.84	0.000

The first test compares the `age` slope for male high school graduates with the `age` slope for males who did not graduate high school. In terms of table 14.2, this is the comparison of β_{2M} with β_{1M}. The difference in these `age` slopes is 175.85 and is significant. The second test is the same as the first test, except the comparison is made for females, comparing β_{2F} with β_{1F}. The difference is 95.54 and is significant. The third and fourth tests compare college graduates with high school graduates. The third test forms this comparison among males and is significant, and the fourth test forms this comparison among females and is also significant.

14.5 Partial interaction on education for the age slope

The three-way interaction can be dissected by forming contrasts on the three-level categorical variable. Say that we use reverse adjacent group comparisons on education, which compares high school graduates with non–high school graduates and college graduates with high school graduates. We can interact that contrast with gender and age, as shown in the `margins` command below.