Health Services Research and Program Evaluation

Causal Inference and Estimation

MARCELO COCA PERRAILLON, RICHARD C. LINDROOTH, DONALD HEDEKER

Draft

Preface

This is a book on quantitative methods in health services research, health economics, and health policy evaluation – more generally referred to as "program evaluation." Health services research is a multidisciplinary field that examines the use, costs, quality, outcomes, and other aspects of health care including the organization of health care markets. Evaluating the impact of health policy is central to the field.

Quantitative analyses in health services research apply methods and language developed in econometrics and statistics/biostatistics. In most applications, the goal is to understand the causal impact of policy changes or "treatments," broadly defined, on a set of outcomes. In most circumstances, however, randomized trials are either not feasible or prohibitively expensive, and we must establish causality using observational data; that is, data that were not collected as part of an experiment. A key distinction between experiments and observational studies is than in observational studies treatment assignment is not under the control of the investigator.

Most readers have already learned that correlation or association does not imply causation. The goal of causal inference is to understand under which conditions correlation –or any other measure of association– does imply a causal effect. Thus, this book is about the design of observational studies and the estimation of statistical models to answer causal research questions. We also cover the necessary background material to understand advanced methods. The background material is focused on understanding the mechanics and properties of parametric and nonparametric statistical models. These models are useful as descriptive and predictive tools, but our ultimate goal is to use them to answer causal research questions.

One feature of our book is that we separate the design of an observational study from the estimation of statistical models. The separation of design and estimation is one of the most valuable aspects of the potential outcomes framework since causal effects are defined independently of an estimation method. This approach is part of the "new" causal inference field in statistics, although causal inference has always been central to econometrics. In the last two to three decades, these separate but related fields have found plenty of common ground regarding causality. The new part is a clear definition of causal effects and a mathematical notation based on potential outcomes and counterfactuals that continues to expand and clarify our understanding of established methods and facilitates the development of new ones.

Our approach is based on the premise that complex concepts are better understood when first introduced with intuitive examples and graphs, followed by theory, and then practical applications using statistical software. Based on our experience teaching graduate-level classes, we think that students learn best by doing, and "doing" means relating the theory

to application using statistical software. Some concepts are difficult to understand in theory but are relatively easy to understand when implemented in practice (and vice versa).

We strive to present theory intuitively but formally to show *how* the theory is applied and *why* methods work, which is essential for understanding *when* specific methods should be used and *what* meaning can be derived from the estimators. This is not a "cookbook approach" book in the sense that we do not focus on rules for specific situations. We do not shy away from presenting complex concepts and mathematical notation because they are essential tools to develop intuition on how and why statistical methods work. Mathematics is a language that makes the job *easier*, not more difficult. Mathematics allows us to represent ideas and concepts using symbols, and we manipulate these symbols to discover new ideas and prove propositions that might not be self-evident. Manipulating complex ideas in our minds without the use of symbols is much more difficult. However, we always provide the intuition behind the mathematics to help students understand how the symbols relate to ideas since not all students are comfortable with mathematics. At the end of the course(s), students should be able to understand the language of mathematics as it applies to statistical analysis.

This book is intended for advanced undergraduates, master's students, and doctoral students in health services research, health economics, public policy, and related fields. Students in these disciplines come from diverse backgrounds with different levels of preparation. We assume the same background that is commonly required for admission to these programs: two semesters of calculus and introductory statistics. A class on linear regression would be helpful, but not strictly necessary since we review the essential features of linear models. We keep linear algebra to a minimum. The goal of the mathematical appendix is to review the mathematical background needed to understand the rest of the book. We hope that students go over the introductory material even if it is not assigned by instructors. Each new concept is based on previous concepts; it is a lack of knowledge of the basics, and the corresponding notation, that confuses students the most. Previous knowledge of Stata is helpful, although the background chapters also serve as an introduction to Stata.

Key features of this book include:

- *Semantics Boxes* that clarify how terms are used in different disciplines. Because our field is multidisciplinary, the terms we use can be confusing –sometimes comically so– because the same terms can have different definitions or because the same concept is named differently in other fields.
- *Notation Boxes* that clarify how mathematical symbols are used in different disciplines or by different authors. As we said, mathematics is a language, but it is a language with symbols that are not standard and can be defined in different ways by different authors. We clarify and present alternative mathematical notation because not understanding unfamiliar notation can prevent students and practitioners from grasping the underlying concepts. A variant of this theme is that sometimes the notation is the result of giving statistical models an interpretation tied to an underlying theory, so we also cover different ways of understanding and/or deriving statistical models. We think students will be better equipped to understand theoretical papers and more advanced textbooks if they understand the notation.

- Extensive examples using datasets to illustrate real-life applications. One frustrating aspect of teaching health services research methods is that we usually cannot use the same datasets that are common in the field and our own research because Data Use Agreements do not permit the distribution of these data. However, we have created multiple datasets from publicly available sources and include datasets that authors have made publicly available to reproduce published papers. Our goal is to use datasets that reflect how practitioners work in our fields.
- Stata code to reproduce all examples and figures in the book. We use Stata code as a tool for learning. In some cases, like graphs or long output, not all of the code is in the book, but it is available in the online supplemental material.
- Stata version control. We prefer Stata because it has the features we need and it has extensive documentation and substantial technical support. Another key feature of Stata is that it is backwards compatible. Regardless of updates, commands will always work provided the code includes a Stata version statement. This ensures that our code will not become obsolete when new versions are released or commands are updated. Most of our code requires Stata 16.1, but some examples require Stata 17. Each program file begins with a version statement.
- Online supplemental material. The online supplemental material includes R code to replicate most of the examples in the book when possible, although some material is specific to Stata. The online supplemental material also covers additional topics that we had to leave out from the text because of space constraints.
- End-of-chapter exercises to reinforce key concepts.
- End-of-chapter bibliographical notes with references to books and papers where readers can find additional or complementary material.

This book is also intended to be a tool for faculty who teach quantitative methods and a reference for practitioners. We wrote it because we could not find a textbook that fit the needs of students. In our classes, we ended up assigning book chapters and papers that use different notation and language, which makes both learning and teaching more difficult. We had to complement those materials with extensive lecture notes and "translations" of notation, terms, and subject-matter. Our lecture notes are the basis for this book.

Additional supplemental material for instructors include:

- Solutions to end-of-chapter exercises.
- Most of the sample datasets contain additional variables that are not part of our analyses. Instructors could use these variables to expand problems sets or create examples focusing on different research questions. In many cases, the variables have missing values. Most textbooks use small sample datasets with non-missing values, but this does not reflect the reality of how research is conducted, so we decided to retain missing values in some of the datasets.
- Lecture notes for most chapters. The lecture notes focus on the most important parts of each chapter. These notes can be used as a starting point for teaching with our book.
- Errata. Despite multiple revisions and editing, the presence of a mistake converges to 1 in probability given the length of our book. We will post a complete list of errors by chapter as we find them, including updates and clarification of some material.

We wrote the book with a two-semester quantitative methods sequence in mind plus additional material for review. We cover topics that should be the standard toolkit in health services research and health/public policy doctoral programs as well as applied econometrics courses in economics programs, although most of our examples are about health care.

The book is divided into four parts. Parts I and II introduce the major subjects we cover, including the potential outcomes framework and a review of statistical concepts and linear regression. Part III focuses on estimation and inference of statistical models, including interpretation of model parameters (causal or not) and discussion of nonparametric models. In other words, Part III discusses techniques to estimate statistical models and the assumptions and properties of these models when applied to a sample, *without assuming that findings from these models have a causal interpretation*. On the other hand, Part IV covers the most important methods to estimate causal effects using observational data: propensity scores and matching estimators as an alternative and complement to regression adjustment, longitudinal (panel) data, difference-in-differences, regression discontinuity designs, and instrumental variables.

Two chapters are fundamental for students to master: Chapter 3 on the potential outcomes framework and Chapter 6 on marginal effects. Chapter 3 is the foundation to understand the definition of causal effects and the identification of causal effects using a sample, and it presents the potential outcome notation we use in the rest of the book. Chapter 6 on marginal effects is essential for understanding the interpretation of model parameters and to express model parameters in different metrics regardless of whether the parameters have a causal interpretation. We provide an overview of each chapter and their connections in Chapter 1.

We have tried to make the chapters as self-contained (modular) as possible –particularly in Part IV– so they can be used independently, although this separation is artificial. We refer to other material in the book when we think students would benefit from reading sections in other chapters, but we have tried to keep such references to a minimum. Each chapter progresses from simple to advanced, from known to unknown, and from concrete to abstract without losing track of practical applications. Instructors could skip the sections that appear towards the end of each chapter if they think the material is too advanced for their students. However, we hope that all of the material can be covered, time permitting. Often, "advanced" really means "unknown." Most concepts are simple once we understand them, and our understanding of "sophisticated" changes with time. What was a sophisticated method a decade ago could be a standard one now.

A typical two-semester sequence for students starting a sequence of quantitative methods would cover:

First semester:	Chapters 3-6
Second semester:	Chapters 9-13
Optional topics:	Appendix, Chapters 2, 7 and 8

In some programs, students take a year of mathematical statistics and/or econometric theory before taking applied methods classes. In this case, a two-semester sequence would skip some of the background material but cover additional chapters:

First semester: Chapters 3, 4-7
Second semester: Chapters 9-13
Optional: Chapter 8

Alternatively, the book could be used for a one-semester class on causal methods for the analysis of observational data assuming the statistical/econometrics background material is known:

Chapters 3, 9-13
Optional (but strongly suggested): Chapter 6

Box 4.1 **Notation: Writing probit and logit models**

There are many ways of writing logit and probit models, with different combination of notation that might cause unnecessary confusion. In the latent approach, we often write (omitting subscripts) $\Pr(y = 1|x_1, ..., x_p) = F(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$ or more compact, $\Pr(y = 1|\boldsymbol{x}) = F(\boldsymbol{x'\beta})$ using vectors or even $\Pr(y|\boldsymbol{x}) = F(\boldsymbol{x'\beta})$. In a model with few covariates, we could also write $\Pr(y = 1|x_1, x_2)$. As we saw in Chapter 4, we have many choices for $F(\cdot)$. By convention, $\Pr(y = 1|\boldsymbol{x}) = \Lambda(\boldsymbol{x'\beta})$ denotes a logit model (Λ is capital lambda; the standard logistic CDF). A probit model is often written as $\Pr(y = 1|\boldsymbol{x}) = \Phi(\boldsymbol{x'\beta})$ (Φ is capital phi, the standard normal CDF). Since y is an indicator or dummy variable, then $E[y] = \Pr(y = 1)$. Therefore, we could also write the models above as $E(y|\boldsymbol{x}) = F(\boldsymbol{x'\beta})$. Subscripts are important since they communicate data structure. We could write $\Pr(y_i = 1|\boldsymbol{x}_i) = F(\boldsymbol{x}_i'\boldsymbol{\beta})$, although it is clear from the context in this chapter that we are working with **cross-sectional** data. In Chapter 7 on choice models, we could have j "choices" or observations by person i, so we could write: $\Pr(y_i = j|\boldsymbol{x}_i,) = F(\boldsymbol{x}_i'\boldsymbol{\beta})$ if the choice only depends on the attributes of the decision-maker and not the choice. If the attributes of the alternative choices were of interest, we could include these with a subscript j: $\boldsymbol{x}_{ij}'$. Our preference is to use the notation that matches the scale of estimation, which means matching Stata or R output. For this reason, we write a logit model as $log(\frac{p_i}{1-p_i}) = \boldsymbol{x}_i'\boldsymbol{\beta}$. Note that this is the same as $logit(p) = \boldsymbol{x'\beta}$ or $log[\frac{\Pr(y_i=1|\boldsymbol{x})}{\Pr(y_i=0|\boldsymbol{x})}] = \boldsymbol{x}_i'\boldsymbol{\beta}$. For the probit model, the scale of estimation matches $\Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$. Be careful with look alike Greek letters: another source of confusion is between capital phi, $\Phi(\cdot)$, the standard normal *cumulative* density function (CDF), and lower-case phi $\phi(\cdot)$, the standard normal probability *density* function (PDF). We chose to denote probability as $\Pr(\cdot)$, but all the expressions above could have used $P(\cdot)$ instead.

6 Marginal effects to interpret regression parameters

Marginal effects are used to interpret regression parameters. They can be computed as "what if" predictions of model outcomes under different scenarios. Model predictions quantify the impact of changing the value of a covariate of interest. More technically, and in most models, the marginal effect of a continuous covariate is the numerical partial derivative of the expected value of the outcome with respect to a small change in the covariate. For a discrete covariate, they are measured as incremental changes in expected values of the outcome given a discrete change in the covariate.

Marginal effects are indispensable to interpret parameters in nonlinear models. As we saw in Chapter 4, the scale of estimation in nonlinear models is often different than the scale of interest. In logit models, we are interested in understanding the effect of a covariate in the probability scale, but models are estimated in the log-odds scale. In Poisson models, we are interested in the effect of covariates on the number of events (counts) or the rate of the event, but Poisson models are estimated in log counts or log rates. Marginal effects allow us to report results in various scales and metrics, regardless of whether the parameters have a causal interpretation.

In this chapter, we focus on logit and probit models because marginal effects are often introduced and motivated in the context of these models, but we present general algorithms and examples using other statistical models, including nonparametric models. We first motivate the need for marginal effects with simple examples in which *analytical* partial derivatives can be easily calculated and interpreted. To understand how *numerical* partial derivatives are calculated, we review basic calculus concepts and present algorithms to compute numerical partial derivatives, which involve making decisions on how to hold constant the values of other covariates. We present multiple examples using the -margins- command to compute marginal effects. Because marginal effects are based on predictions, the -margins- command can also be used to obtain adjusted predictions and plots. We also show how standard errors for marginal effects are calculated.

Besides understanding the conceptual definition and practical computation of marginal effects, an important takeaway of this chapter is that in some scales the impact of covariates on outcomes is no longer additive and separable.

6.1 Why do we need marginal effects?

As we discussed in Chapter 2, in simple linear/OLS models, interpreting model parameters is usually straightforward. For example, we showed that the model $y_i = \beta_0 + \beta_1 age +$

Box 6.1 Semantics: Different meanings of "marginal"

The term **marginal effect** causes plenty of confusion in interdisciplinary collaborations. The term emerged from econometrics. In economics, marginal means *additional* or *incremental*. Mathematically, it is a **derivative**. In producer theory, a profit-maximizing firm stops producing a good when its marginal revenue equals its marginal cost; that is, the revenue from an additional quantity is the same as the cost of producing the additional quantity. For a statistician, the first concept that comes to mind after reading marginal is not a derivative but its opposite, a **definite integral**, because of **marginal distributions**. Given the joint probability distribution function of random variables, the marginal probability density function of one variable can be obtained by integrating (or summing with discrete random variables) over the domain of the other variables.

$\beta_2 male + \epsilon_i$, assuming $\epsilon \sim N(0, \sigma^2)$, is a conditional expectation function. We can write the model as $E[y_i|\boldsymbol{x}] = \beta_0 + \beta_1 age + \beta_2 male$. The effect of age on the expected value of the outcome is β_1. Assuming age is measured in years, a one-year increase in age increases the average outcome by β_1. This effect is constant (linear), so the effect of a 10-year increase is $10 \times \beta_1$. A simple linear specification assumes that aging one year has the same impact on teenagers as it does on the elderly. However, even simple linear/OLS models could have non-constant effects, which complicates interpretation. In the model, $y_i = \alpha_0 + \alpha_1 age_i + \alpha_2 age_i^2 + u_i$, the effect of age depends on the value of age: $\frac{dE[y_i|\boldsymbol{x}]}{dage} = \alpha_1 + 2\alpha_2 age$.

With interactions, interpretation is slightly more difficult. In the model $y_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 age_i \times male_i + \epsilon_i$, the effect of age is no longer constant (linear), and the coefficient for age, β_1, needs to be interpreted in relation to β_3. In this model, the effect of age depends on whether the person is male or female. As above, calculating derivatives or incremental changes facilitates interpretation: $\frac{\partial E[y_i|\boldsymbol{x}]}{\partial age} = \beta_1 + \beta_3 male$ and $\frac{\Delta E[y_i|\boldsymbol{x}]}{\Delta male} = \beta_2 + \beta_3 age$. Remember that interactions apply in both directions: the effect of a change in age on the average outcome depends on sex, but the previous model also assumes that the incremental effect of sex depends on age.

In other models, especially nonlinear models, interpretation is more difficult. Consider the following logit model: $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 age_i + \beta_2 male_i$. As we saw in Chapter 4, β_1 is the effect of age in the *log-odds scale*. The sign and the p-value of β_1 provide information on the direction of the effect and statistical significance for a null hypothesis $H_0 : \beta_1 = 0$, but expressing results in the log-odds scale is of no practical interest. We could, however, easily express results in terms of odds ratios, e^{β_1} or e^{β_2}, but odds ratios have many drawbacks. One of this drawbacks is that odd ratios tend to be incorrectly interpreted as a ratio of probabilities, but they are not: odds ratios are a ratio of odds. The odds of an event is the probability that the event occurs divided by the probability that the event does not occur. If greater than one, then the event has a larger chance of occurring than not.

It is more useful to interpret logistic models in the *probability* scale because research

questions are usually focused on understanding the influence of covariates on the probability of an event. We can go from log-odds to probabilities using the **inverse logit function**

$$\Pr(y_i = 1|\boldsymbol{x}) = \frac{e^{(\beta_0+\beta_1 age_i+\beta_2 male_i)}}{1+e^{(\beta_0+\beta_1 age_i+\beta_2 male_i)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 age_i+\beta_2 male_i)}}. \tag{6.1}$$

To calculate the effect of age, we could take the derivative $\frac{\partial \Pr(y_i=1|\boldsymbol{x})}{\partial age}$ as with linear models. However, this derivative does not have a simple, easy to interpret expression since the analytic derivative of $f(x) = e^x$ is also $f'(x) = e^x$.

More generally, suppose we estimate the following logit model:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \tag{6.2}$$

which can be rewritten in the probability scale as

$$\Pr(y_i = 1|\boldsymbol{x}) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi})}}. \tag{6.3}$$

The marginal effect for a continuous covariate x_1 is given by the expression:

$$\frac{\partial \Pr(y_i = 1|\boldsymbol{x})}{\partial x_1} = \beta_1 \frac{e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi}}}{(1+e^{-(\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi})})^2} \tag{6.4}$$

Equation 6.4 shows that the effect of an infinitesimal change in x_1 on the probability of the outcome is nonlinear – as it has to be since the outcome must be bounded between 0 and 1. The direction of the change is given by the sign of β_1. It is evident from 6.4 that the effect of x_1 *depends on the value of all other covariates in the model* even if the underlying model does not include interactions. In the probability scale, there are *implicit interactions*. Effects are *no longer additive and separable*. This is an important feature shared with other nonlinear models as we discuss in Section 6.9 and Chapter 7. The implications for modeling are discussed in Section 6.13.

Since working with analytical derivatives in nonlinear models is not as simple as in linear/OLS models, we will calculate numerical derivatives to interpret parameters. The numerical partial derivative is the **marginal effect**. If the variable of interest is discrete, such as an indicator or dummy variable, we want to understand the impact of changing the indicator variable from 0 to 1, which we call the **incremental effect**.

6.1.1 Metrics: Odds ratio, relative risk, risk difference

We first review the **metrics** (or **effect measures**) that can be obtained from logit (or probit) models since we will discuss marginal effects in the context of these models. In particular, the difference between odds ratios, relative risks, and risk differences tends to confuse students even though they are just different ways of quantifying comparisons.

Consider a simple example: the probability of death in a control group is 0.40, while the probability of death in the treatment group is 0.20. This implies an **odds ratio**= $\frac{\frac{0.2}{1-0.2}}{\frac{0.4}{1-0.4}}$ =

0.375 or that treatment reduces the odds of death by a factor of 0.375. Inversely, the odds of death are 2.67 higher in the control group ($\frac{1}{0.375}$). But neither metric is a relative risk (relative probabilities), even though that is how odds ratio is sometimes incorrectly interpreted. The **relative risk** is $\frac{0.2}{0.4} = 0.5$, which means that the *probability* of death is reduced by half in the treatment group (or the probability of death is 2 times higher in the control group). The **risk difference** is the difference between the probabilities: -0.2 (0.4-0.2), implying a 20 percentage point decline.

Something else is critically important: odds ratios and relative risks do not provide a good sense of the magnitude of the relationship. Assume now that the probability of death in the control group is 0.00004 and 0.00002 in the treatment group. The odds-ratio is still 0.375 and the relative risk is still 0.5. However, the risk difference which is how we will express marginal and incremental effects, gives a better sense of the magnitude: $0.00002 - 0.00004 = -0.00002$ or -0.002 percentage points. The practical implications are very different. Think about this example next time you read that a treatment increases the chances of something else happening by a factor of X, followed by a statement about statistical significance. Are they referring to odds or probabilities? Is the practical implication consistent with what the headline suggests? The practical significance of the finding is better understood in the context of the difference in probabilities or with knowledge of the underlying chance of the outcome.

In section 6.7.2, we will use logit and probit models to express results in these metrics, but they are not the only possibilities. For example, we could also use percent change: the treatment reduces the probability of death by 50%: (0.2-0.4)/0.4. Note that the percent change is the relative risk minus 1 multiplied by 100: $\frac{P_t - P_c}{P_c} = (\frac{P_t}{P_c} - 1) \times 100$, where P_c is the probability of the outcome in the control group and P_t is the probability in the treatment group. In applied work, it is useful to communicate results in different metrics.

6.2 Analytical and numerical derivatives

In this section, we review the definition of analytical derivatives because numerical derivatives closely follow the analytical definition. In the next section, we present an algorithm that applies the analytical definition to obtain a numerical approximation using model predictions and observed data.

6.2.1 One-sided derivative

To understand how to calculate numerical derivatives, it is helpful to review basic calculus. An **analytical derivative** is a rate of change; it measures the slope of the tangent line that touches a curve at some point x_0 in the curve. The slope of a line is $m = \frac{y_1 - y_0}{x_1 - x_0} = \frac{\Delta y}{\Delta x}$, that is, the change in y relative to the change in x. The derivative follows the same concept, except that the change around a point x_0 is extremely small (infinitesimal). It is the change in y when the change in x at $x = x_0$ tends to zero. It is a limit:

$$f'(x = x_0) \equiv \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{6.5}$$

If you look at the numerator closely, $f(x_0 + h) - f(x_0)$ is the change in y; that is, Δy. The denominator is the change in x: $\Delta x = (x_0 + h) - x_0 = h$. If the function had other variables, we would consider them as fixed or constant when evaluating the *partial* impact when x changes. All the formulas for the derivative of a function can be derived using the definition above by taking the limit of the change in x (h in Equation 6.5), although shortcuts and rules like the product rule or the power rule simplify derivations. For example, recall that the derivative of $f(x) = x^2 + a$ is $f'(x) = 2x$, because:

$$\begin{aligned} f'(x = x_0) = \lim_{h \to 0} \frac{(x_0 + h)^2 + a - x_0^2 - a}{h} &= \frac{x_0^2 + 2x_0h + h^2 - x_0^2}{h} \\ &= \frac{2x_0h + h^2}{h} \\ &= 2x_0 + h \\ &= 2x_0 \end{aligned} \tag{6.6}$$

For a function to be **differentiable**, it must be continuous or "smooth" at every point x_0. If a function has a discontinuity at some point x_0, we could still analyze the change in y by changing h in one direction, by either taking a right-hand limit ($\lim_{h \to 0^+}$) or a left-hand limit ($\lim_{h \to 0^-}$). A function is differentiable at x_0 if the right-hand side and left-hand side derivatives exist and are equal.

We could also use the definition of the derivative to calculate the **numerical derivative** of a function. Using Equation 6.5, we plug in a *small* number for h to approximate the derivative at a point x_0. We would have:

$$f'(x = x_0) \approx \frac{f(x_0 + h) - f(x_0)}{h},$$

which is just Equation 6.5 evaluated at h instead of the limit of h.

Consider a simple example to provide some intuition. Suppose $f(x) = x^2$ and $x_0 = 2$. Using $h = 0.0001$, the numerical derivative is

$$f'(x = 2) \approx \frac{(2 + 0.0001)^2 - 2^2}{0.0001} = 4.0001, \tag{6.7}$$

which is close to the actual derivative of 4: $f'(x = 2) = 2x = 2 \times 2 = 4$. Keep in mind that this is a one-sided computation. We increased x at x_0 by a small amount, but we could have decreased x as well (subtracting by 0.0001, you should get 3.9999).

Computationally, it is not trivial to come up with a number h that is "small but large enough" so that the computations are numerically accurate. The examples we cover in this chapter are not as simple as Equation 6.7, where we do not lose much numerical precision. Here, we know how much we lose (because we know the analytic derivative at $x_0 = 2$ is 4): $(4 - 4.001) = 0.0001 = h$. We could reduce the loss by making h smaller, but in general,

we do not want to make it too small either because we could lose numerical precision. We discuss numerical precision in section 6.14.1.

6.2.2 Two-sided, centered derivative

There is another way of defining a derivative, and this is the definition Stata uses to estimate marginal effects because it has computational advantages. A **two-sided derivative** is calculated using both a positive and negative small change h around a point x_0:

$$f_2'(x = x_0) \approx \frac{f(x_0 + h) - f(x_0) - [f(x_0 - h) - f(x_0)]}{2h} = \frac{f(x_0 + h) - f(x_0 - h)}{2h} \tag{6.8}$$

As with a one-sided derivative, the choice of a small h is still an important consideration, but we highlight the advantage of this definition with an example. If we use 6.8 to evaluate the derivative of $f(x) = x^2$ at $x_0 = 2$ with $h = 0.0001$, we would find that the numerical derivative is exactly 4, which matches the analytical partial derivative evaluated at $x_0 = 2$.

In the following examples, we first calculate marginal effects "by hand" and then compare our results to those obtained with the -margins- command. Our calculations will be very close, but Stata also implements an additional iterative procedure that changes the value of h to achieve better numerical accuracy (Section 6.14.1). This can slow down the calculation of marginal effects in some models.

6.3 Average marginal effects

In the previous section, we reviewed the definition of an analytical partial derivative and evaluated the derivative at a particular point numerically by choosing a value for h. In this section, we present an algorithm that applies the same ideas to obtain partial numerical derivatives using a logit regression model.

The algorithm consists of five step and computes the the partial numerical derivative for a small change in x_1, assuming x_1 is a continuous variable:

1. Estimate the logit model $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$.
2. Increase the value of the variable x_1 by a "small" amount h: $x_1 = x_1 + h$. h depends on the units of x_1: $h = (|\bar{x_1}| + 0.0001) \times 0.0001$, where $\bar{x}_1$ is the mean of x_1. For each observation i, calculate predictions $\hat{y}_{1i}$ in the probability scale keeping all other covariate values $(x_{2i}, \ldots, x_{pi})$ as observed.
3. Decrease the value of the variable x_1 by the same small amount h for each observation i. Calculate predictions $\hat{y}_{0i}$ in the probability scale using values for all covariates as observed
4. For each observation i, calculate the difference of the two predictions divided by $2h$: $(\hat{y}_{i1} - \hat{y}_{i0})/2h$
5. The average of this difference is the numerical derivative: $E[\frac{\hat{y}_{1i} - \hat{y}_{0i}}{2h}] \approx \frac{\partial \Pr(y_i = 1|\boldsymbol{x};\beta)}{\partial x_1}$

Conceptually, this procedure follows the two-sided numerical derivative (Equation 6.8) and predictions from the logit model correspond to Equation 6.4. The small change $h = (|\bar{x}_1| + 0.0001) \times 0.0001$ is the same small initial change used by Stata (0.0001 is the same as 10^{-4} in scientific notation; in computer language, it is `1e-4`). Since the unit of measurement of a variable matters (e.g., age in years versus age in decades), it makes sense that the "small" change depends on the mean of the covariate, $\bar{x}_1$. These predictions can be conceptualized as "counterfactual" predictions, even at the risk of confusing this approach with the potential outcomes framework of Chapter 3. However, it is important to understand that we use the *estimated* model to make "what if" predictions by changing the value of a covariate. Said in other words, we use the estimated model to understand *implications* of the model in a scale that is different from the estimation scale.

This algorithm is general, in the sense that the same procedure could be applied to other regression models. We simply estimate a different model in step 1 and calculate the corresponding predictions in steps 2 and 3 using the model estimates.

To show how to implement the algorithm, we use an extract from the 2016-2017 National Health and Nutrition Examination Survey (NHANES). The outcome of interest is the probability that a respondent has high cholesterol (hyperlipidemia), defined as total blood cholesterol higher than 200 milligrams per deciliter. We are interested in understanding how age affects the probability of high cholesterol. Other covariates of interest are sex and and indicator variable equal to 1 if the respondent has any health insurance coverage. Besides changes in diet, medications are effective in treating this condition. We do not observe whether a person is taking medications to lower cholesterol in this sample. However, health insurance coverage could be a proxy for access to medications, as these drugs require a prescription and are expensive. Below is a description of the data:

```
use "nhanes_cholesterol",clear
describe
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
highchol        float   %9.0g                 1 if total cholesterol > 200 mg/dL
age             double  %12.0g                Age: continuous
male            double  %12.0g                1 if male
insured         float   %9.0g                 1 if any insurance

summarize highchol age male insured
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    highchol |      5,214    .3214423    .4670749          0          1
         age |      5,214    48.44208    19.57075         16         80
        male |      5,214    .4854239    .4998354          0          1
     insured |      5,214    .8632528    .3436132          0          1
lpoly highchol age, deg(2) jitter(3)
```

Figure 6.1 shows the relationship between age and the probability of high cholesterol using a **kernel-weighted local polynomial regression**. The graph shows that the expected value of high cholesterol (i.e., probability) conditional on age increases at a steady rate until about age 55, when it starts to decline. Note that children and teens can have high choles-

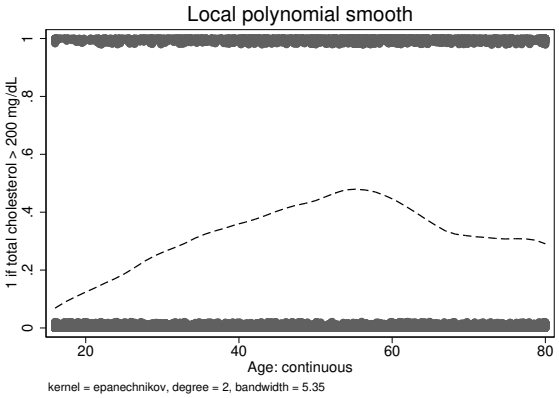

Fig. 6.1 Probability of high cholesterol and age

terol, although the upper-left corner points show that there are fewer teen respondents with high cholesterol.

To implement the algorithm, we first estimate the following logit model:

$$log\left(\frac{highchol_i}{1-highchol_i}\right) = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 insured_i \tag{6.9}$$

Next, we implement steps 2 to 5 to calculate the marginal effect of age on the probability of high cholesterol:

```
1  preserve
2     quietly summarize age
3     scalar h = (abs(r(mean))+0.0001)*0.0001
4     quietly logit highchol age male insured
5     clonevar age_c = age
6     replace age = age_c + scalar(h)
7     predict double hc_1 if e(sample)
8     replace age = age_c - scalar(h)
9     predict double hc_0 if e(sample)
10    generate double dydx = (hc_1-hc_0)/(2*scalar(h))
11    summarize dydx
12 restore
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      5,214    .0036087    .0004465   .0025876   .0042398
```

We used `preserve` to save the current data in memory because in this example we do not want to keep the variables created after the preserve command. The command `restore` restores the dataset to the version before the command `preserve` was executed. Since we need to change the value of `age` twice, we use the command `clonevar` in line 5 to create a replica of `age`. By default, the command `predict` calculates predictions from a logit model in the probability scale. Note that the variable `dydx` is the difference in predictions (in the probability scale) divided by the small change for *each observation*. The **average marginal**

effect is the mean of the individual marginal effects: .0036087. The interpretation is that a small increase in age increases the probability of high cholesterol by 0.36 percentage points, holding other covariates constant.

We now use the -margins- command to replicate our "manual" calculations. We first estimate the same logit model using the `quietly` prefix since we are not interested in seeing the output of the logit command.

```
quietly logit highchol age male insured
margins, dydx(age)

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0036087   .0003255    11.09   0.000     .0029708    .0042467
------------------------------------------------------------------------------
```

In this case, the -margins- output matches the manual calculation of marginal effects for a continuous variable (.0036087), although they may not always exactly match; the values could be off by some decimals. The reason is that Stata uses an algorithm to ensure numerical precision (see Section 6.14.1). Note that Stata also calculates the standard error and confidence intervals for the marginal effect, which are usually not the same as the standard error of the corresponding model coefficient. We explain how to estimate standard errors for marginal and incremental effects in Section 6.12.

6.4 Average incremental effects

The discussion so far focused on marginal effects for continuous variables. Calculating incremental effects for indicator variables is simpler, although it follows a similar conceptual idea. We want to understand model implications when an indicator or dummy variable changes from 0 to 1 – female to male; uninsured to insured. Because indicator variables are commonly coded as 0 or 1, the denominator of the rate of change is 1. Below, we calculate the incremental effect for male:

```
preserve
   quietly logit highchol age male insured
   replace male = 0
   predict double hc_0 if e(sample)
   replace male = 1
   predict double hc_1 if e(sample)
   generate double dydx = (hc_1-hc_0)
   summarize dydx
restore
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      5,214   -.0504936    .0059993  -.0591639  -.0389134
```

The incremental effect is -.0504936, which means that a male respondent is 5.04 percentage points less likely to have high cholesterol than a female respondent, holding age and insurance status constant.

Using the -margins- command with the `dydx(male)` option, we can replicate our manual calculation of the incremental effect:

```
quietly logit highchol age i.male i.insured
margins, dydx(male)

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : 1.male

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.male |  -.0504936   .0127817    -3.95   0.000    -.0755452   -.0254419
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

In contrast to the marginal effect of a continuous variable, the manual calculation and the -margins- command of incremental effects will exactly match because there is no need to fine-tune the infinitesimal change h in the covariate. Be mindful of the note at the end of the output:

```
Note: dy/dx for factor levels is the discrete change from the base level.
```

This is Stata's way of letting us know that it considered the variable `male` to be an indicator variable. Stata knows this because the model we estimated used the factor variable operator `i`. If we had omitted the operator, Stata would have assumed `male` to be a continuous variable. Instead of an incremental effect, we would have obtained the marginal effect by changing `male` by a small amount, as though it were continuous. When using the -margins- command, always use the factor variable operator when estimating models.

We can also obtain the marginal or incremental effect for all variables at once using the wildcard operator `*`:

```
quietly logit highchol age i.male i.insured
margins, dydx(*)

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age 1.male 1.insured

------------------------------------------------------------------------------
             |            Delta-method
```

Box 6.2 Semantics: Marginal and incremental effects and other terms

In this chapter, we use two separate terms depending on whether a variable is continuous or discrete. **Marginal effects** refers to the effect of a continuous variables and are conceptualized as two-sided numerical derivatives. **Incremental effects** refers to the effect of discrete variables, conceptualized as the change in the outcome when the indicator variable switches from 0 to 1. However, we often use the term "marginal effects" to refer to both, although we make the distinction when needed. Another name for marginal effects is **average partial effects**, which includes effects of continuous and discrete covariates. In the context of logit or probit models, marginal effects are also called **average predicted probabilities**.

```
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
         age |   .0036087   .0003255    11.09   0.000     .0029708    .0042467
      1.male |  -.0504936   .0127817    -3.95   0.000    -.0755452   -.0254419
   1.insured |  -.0371851   .0196002    -1.90   0.058    -.0756008    .0012307
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Holding age and sex constant, insured individuals are 3.71 percentage points less likely to have high cholesterol.

6.5 Holding constant the value of other covariates

One consequence of computing numerical marginal and incremental effects using estimated models and data is that we need to decide how to fix or hold constant the values of the other covariates. In the preceding examples, we left the value of the other covariates at their observed values. In steps 2 and 3 of the algorithm to obtain marginal effects (Section 6.3), predictions were calculated for each observation i; in our example, each NHANES respondent.

With analytical derivatives, other covariates become constants, but numerical derivatives involve model predictions, and it is clear from Equation 6.1 that predictions will change if we used different values for covariates. We have several options. We could leave covariates at their observed values (as we did so far) or hold them at their means or at representative values that could be of scientific interest. For dummy variables, holding covariates at their means implies using a value that is, for example, the proportion of females in the sample if the variable is an indicator for female.

In practice, the choice of values for the other covariates usually does not matter when calculating marginal or incremental effects, the *difference* between predictions. It is easier to understand the impact of other covariates using simple linear/OLS model. In Step 4

of the algorithm, we took the difference between two predictions for *each* observation i: $(\hat{y}_{i1} - \hat{y}_{i0})/2h = \frac{1}{2h}(\hat{y}_{i1} - \hat{y}_{i0})$. Suppose, for example, that we had estimate the model $y_i = \beta_0 + \beta_1 age_i + \beta_3 x_i + \epsilon_i$ and want to calculate the marginal effect of age. Then step 4 of the algorithm for an observation i would be:

$$\begin{aligned}\frac{1}{2h}(\hat{y}_{i1} - \hat{y}_{i0}) &= \frac{1}{2h}[\hat{\beta}_0 + \hat{\beta}_1(age_i + h) + \hat{\beta}_3 x_i - (\hat{\beta}_0 + \hat{\beta}_1(age_i - h) + \hat{\beta}_3 x_i)] \\ &= \frac{1}{2h}[\hat{\beta}_1(age_i + h) - \hat{\beta}_1(age_i - h)] \\ &= \hat{\beta}_1\end{aligned} \tag{6.10}$$

The expression 6.10 shows that the covariate x will cancel out, as would any other covariate in the model in this simple example with no interactions. Although we use the values of covariates as they were observed, we use the same values when calculating the "what if" prediction for each observation i. Therefore, as long as the same values for the covariates are used to calculate the predictions, the other covariates and their estimated coefficients will cancel out when taking the difference.

In logit (and other nonlinear models), there could be a difference, although it tends to be small in practice. The difference is due to the fact that the other covariates do not cancel when taking the difference out because of the non-linearity induced by the (inverse) logit function (Equation 6.1). There is no general rule that tells us the direction or the magnitude of the difference, only that in practice the difference tends to be small (Chapter 17 in Greene, 2018 for an argument using second order Taylor series).

We interpret average marginal and incremental effects as "holding other covariates constant" because for each observation i we are not changing the value of the other covariates when calculating marginal effects, even if by default Stata does not hold other covariates constant at the same values for *all* observations. However, we could hold covariates constant at different values for all observations. For example, the marginal effect of age holding male and insured at their sample mean is:

```
quietly logit highchol age i.male i.insured
margins, dydx(age) at((mean) male insured)

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
at           : 0.male          =    .5145761 (mean)
               1.male          =    .4854239 (mean)
               0.insured       =    .1367472 (mean)
               1.insured       =    .8632528 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0036125   .0003253    11.10   0.000     .0029749    .0042501
------------------------------------------------------------------------------
```

A small increase in age increases the probability of high cholesterol by 0.36 percentage points, holding male and insurance status at their means. This is almost identical to the result we obtained (.0036087) when we left sex and insured as they were observed.

It is always a good idea to carefully read the -margins- command output. Stata tells us that the numerical derivative was with respect to `age` (`dy/dx w.r.t. : age`) at specific values of `male` and `insured`, and it shows the specific values it uses. The example above is not the same as `margins, dydx(age) atmeans` because Stata would also calculate the numerical derivative at the mean of `age`.

We could also evaluate marginal effects holding other covariates at specific values that could be of scientific interest. For example, it may be of interest to calculate the marginal effect of age for males who do not have health insurance:

```
margins, dydx(age) at(male=(1) insured=(0))

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
at           : male            =           1
               insured         =           0

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .003644   .0003561    10.23   0.000      .002946    .0043419
------------------------------------------------------------------------------
```

As before, we do not see a large change when we hold covariates at specific values. The code below replicates the above -margins- result:

```
1   preserve
2      quietly summarize age
3      scalar h = (abs(r(mean))+0.0001)*0.0001
4      quietly logit highchol age male insured
5      clonevar age_c = age
6      replace male = 1
7      replace insured = 0
8      replace age = age_c + scalar(h)
9      predict double hc_1 if e(sample)
10     replace age = age_c - scalar(h)
11     predict double hc_0 if e(sample)
12     generate double dydx = (hc_1-hc_0)/(2*scalar(h))
13     summarize dydx
14  restore
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |      5,214     .003644    .0004217   .0028666   .0041956
```

In lines 6 and 7, we set the values for *male* = 1 and *insured* = 0, which we used for all the predictions.

Using the `at()` option is also helpful to further understand the implications of nonlinear

models. In the linear/OLS model $y_i = \alpha_0 + \alpha_1 age_i + \alpha_2 male_i$, the effect of age is the same regardless of whether an observation is male or female because the model does not include interaction terms between age and male. We would obtain the same effect for age when male is 1 than when male is 0. However, in the logit model $log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 age_i + \beta_2 male_i$, this is no longer the case in the *probability* scale even if the model does not have explicit interactions between age and male in the log-odds scale (see Equation 6.4). Thus, the marginal effect with respect to age at all combinations of male and insurance yields four different estimates:

```
quietly logit highchol age i.male i.insured, nolog
margins, dydx(age) at(male=(0 1) insured=(0 1)) vsquish
Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM
Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
1._at        : male            =           0
               insured         =           0
2._at        : male            =           0
               insured         =           1
3._at        : male            =           1
               insured         =           0
4._at        : male            =           1
               insured         =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0038945    .000368    10.58   0.000     .0031732    .0046158
          2  |   .0037208   .0003345    11.12   0.000     .0030652    .0043765
          3  |    .003644   .0003561    10.23   0.000      .002946    .0043419
          4  |   .0034227   .0003092    11.07   0.000     .0028167    .0040286
------------------------------------------------------------------------------
```

(The option `vsquish` suppresses extra blank lines so the output is more compact.)

Evaluating numerical derivatives at specific values of covariates is particularly helpful in models with interactions since interacted models assume that the effect of interest depends on the value of other covariate(s). Of course, we could have added interactions in the model above. We discuss interactions in the context of logit models in Sections 6.10 and 6.13.

6.6 Evaluating the derivative at specific values of the variable of interest

In the previous section, we were fixing the value of *other* covariates. However, we could also be interested in evaluating the numerical derivative at specific values of the variable of interest. For example, we may be interested in understanding how the relationship between age and the probability of high cholesterol changes at different ages since Figure 6.1

suggests a nonlinear relationship. In effect, we are now interested in the partial numerical derivative corresponding to the partial analytical derivative

$$\frac{\partial \Pr(highchol = 1|\boldsymbol{x})}{\partial age}\Big|_{age=a}, \tag{6.11}$$

where a is a specific age.

The value at which the marginal effect is evaluated matters because the relationship between the variable of interest and the outcome may change signs or have a strong curvature. Another way to think of this is that the *second* derivative may change sign and magnitude depending on where it is evaluated. As an aside, it is also possible to obtain the second derivative numerically; see (Gould et al., 2006, Chapter 1).

Below, we evaluate the marginal effect of age at different values of age:

```
quietly logit highchol age male insured
margins, dydx(age) at(age=(16 40 55 75 80)) vsquish

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM
Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
1._at        : age             =          16
2._at        : age             =          40
3._at        : age             =          55
4._at        : age             =          75
5._at        : age             =          80
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0028243    .000182    15.52   0.000     .0024676     .003181
          2  |   .0034622   .0003024    11.45   0.000     .0028694     .004055
          3  |   .0038028   .0003657    10.40   0.000      .003086    .0045196
          4  |   .0041194   .0004105    10.04   0.000     .0033149     .004924
          5  |   .0041669   .0004122    10.11   0.000     .0033591    .0049748
------------------------------------------------------------------------------
```

Once again, even though age is linear on the log-odds scale, we obtain different marginal effects (rates of change in the probability scale). The change in the probability of high cholesterol is lower at age 16 (.28 percentage points) than at age 75 (0.42 percentage points), holding sex and insurance status constant. However, the relationship between age and the probability of high cholesterol may be better modeled by explicitly specifying age nonlinearly using a quadratic term, which is statistically significant (not shown here):

```
quietly logit highchol c.age##c.age male insured
margins, dydx(age) at(age=(16 40 55 75 80)) vsquish

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM
Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
```

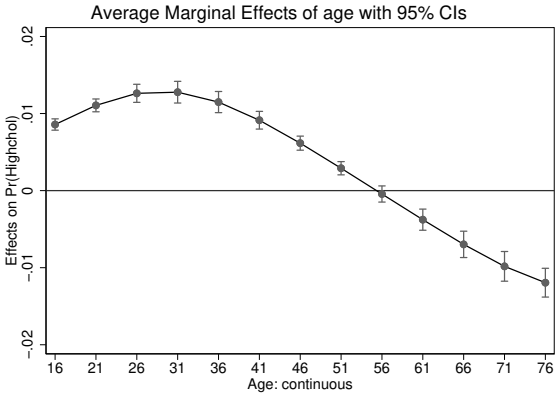

Fig. 6.2 Marginal effect of age evaluated at different ages

```
1._at         : age             =          16
2._at         : age             =          40
3._at         : age             =          55
4._at         : age             =          75
5._at         : age             =          80
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0085765   .0003743    22.91   0.000     .0078429      .00931
          2  |   .0096673   .0006128    15.78   0.000     .0084663    .0108683
          3  |    .000232   .0005074     0.46   0.647    -.0007625    .0012265
          4  |  -.0116031    .000974   -11.91   0.000     -.013512   -.0096941
          5  |  -.0128063   .0007878   -16.26   0.000    -.0143504   -.0112622
------------------------------------------------------------------------------
```

We see the change in *sign* and the peak close to age 55, where the partial derivative is close to zero. The nonlinear relationship between the marginal effect of age and the value of age is apparent in Figure 6.1, which shows that the marginal effect of age is positive in some age ranges and then turns negative, which corresponds to the previous results.

Graphs are always helpful for interpretation. The postestimation command -marginsplot- can be used to graph the output of the -margins- command:

```
quietly {
   logit highchol c.age##c.age male insured
   margins, dydx(age) at(age=(16(5)80)) vsquish
   marginsplot, yline(0)
}
```

Figure 6.2 shows the marginal effect of age on the probability of high cholesterol evaluated at age 16 to 80 by increments of 5. The slope or partial derivative of age turns negative at about age 56.

6.7 Predictive margins

Marginal and incremental effects are based on "what if" model predictions. So far, we took the difference of these predictions, but we may be interested in exploring the predictions themselves. We could use these predictions to calculate different metrics that quantify the effect of variables on the outcome.

For example, when estimating *incremental* effects, we calculated predictions for each observation after male was set to 0 and then again after setting it to 1. The average of these counterfactual predictions yields what Stata calls **predictive margins**. The incremental *effect* of a dummy variable is the difference between the two predictive margins. Step 5 of the algorithm in section 6.3 calculates the average of the marginal effect as $E[\frac{\hat{y}_{1i}-\hat{y}_{0i}}{2h}]$. For incremental effects $h = 1$ so step 5 simplifies to: $E[\hat{y}_{1i}] - E[\hat{y}_{0i}]$; the expected value of the difference is the same as the difference in expected values.

Below, we compute predictive margins for insured:

```
quietly logit highchol c.age##c.age i.male i.insured
margins insured, post

Predictive margins                              Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     insured |
           0 |   .3287234    .017535    18.75   0.000     .2943554    .3630914
           1 |   .3203157   .0067516    47.44   0.000     .3070827    .3335486
------------------------------------------------------------------------------
```

The average predicted probability of high cholesterol if none of the respondents were insured is .3287234, but if all were insured, the predictive probability falls to .3203157. Note that predictive margins are calculated without the `dydx()` option. The `post` option allows us to store the margins estimates in the matrix `e(b)` so they can be used later.

We can verify that the difference between predicted probabilities is identical to the incremental effect (-.0084077):

```
matrix list e(b)
e(b)[1,2]
            0.          1.
      insured    insured
y1   .3287234  .32031569

display  e(b)[1,2] - e(b)[1,1]
-.0084077

quietly logit highchol c.age##c.age i.male i.insured
margins, dydx(insured)

Average marginal effects                        Number of obs     =      5,214
```

```
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : 1.insured

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.insured |  -.0084077   .0188789    -0.45   0.656    -.0454096    .0285942
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Predictive margins are also called **adjusted predictions**. In this example, we controlled for age and sex, although we did not hold these covariates fixed at any specific value. In the context of linear/OLS models, predictive margins are also referred to as **adjusted means**.

6.7.1 Predictive margins for a continuous variable

We cannot calculate predictive margins for continuous variables using the syntax presented in the previous section because a continuous variable takes an infinite number of values. However, we can still use the -margins- command by choosing specific values with the `at` option:

```
quietly logit highchol c.age##c.age i.male i.insured
quietly margins, at(age=(16(1)80) (mean) male insured)
marginsplot, recastci(rarea) plotopts(msymbol(none))
```

Keep in mind that this is the third syntax variation of the -margins- command in this chapter. There is no variable following `margins` and no `dydx()` option. Here, we calculated model predictions for each age 16 to 80 at one-year intervals, and we held male and insured at their sample mean. We used the `quietly` prefix because the output is long, reflecting each age-specific estimate. The -marginsplot- command plots the predictions from -margins- (Figure 6.3). We specified -marginplots- options `recastci()` and `plotopts` to change how the confidence intervals and the line are displayed.

The difference between Figure 6.2 and Figure 6.3 is illuminating. Both depict model implications, but they do so in different ways. Figure 6.2 plots the rates of change (marginal effects) and Figure 6.3 plots adjusted predictions. Note that the highest probability occurs at between age 55-56, which corresponds to the point when the marginal effect in Figure 6.2 becomes negative. We will use adjusted plots in subsequent chapters, especially in Chapter 11 to graph adjusted difference-in-difference models.

6.7.2 Odds ratios, relative risks, and risk differences redux

We now return to the discussion of Section 6.1.1 to show that odds ratios, relative risks, and risk differences can also be calculated using predictive margins. We use the `post` option to store predictive margins in matrix `e(b)` and then calculate relative risks and odds ratios:

```
quietly logit highchol c.age##c.age i.male i.insured
quietly margins male, post
```

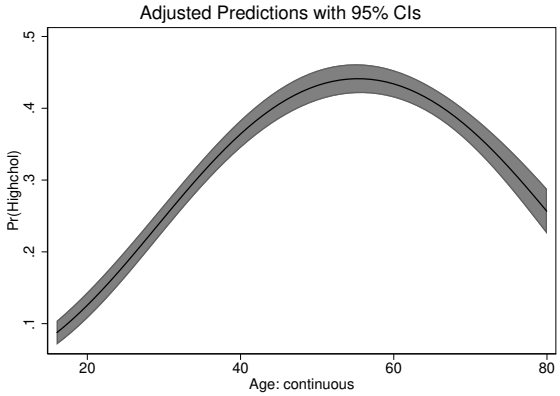

Fig. 6.3 Probability of high cholesterol by age, holding sex and insurance status at sample means

```
matrix list e(b)
e(b)[1,2]
            0.          1.
         male        male
y1  .34410891  .29744087

* Relative risk
display e(b)[1,2] / e(b)[1,1]
.86438004

* Odds ratio
display (e(b)[1,2]/(1-e(b)[1,2]) )/ (e(b)[1,1]/(1-e(b)[1,1]))
.80696293
```

For reference, here is the model with all the coefficients in the odds scale:

```
logit highchol c.age##c.age i.male i.insured, nolog or
Logistic regression                             Number of obs     =      5,214
                                                LR chi2(4)        =     341.11
                                                Prob > chi2       =     0.0000
Log likelihood = -3103.5878                     Pseudo R2         =     0.0521
------------------------------------------------------------------------------
    highchol | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.162948   .0115219    15.24   0.000     1.140584    1.185751
             |
 c.age#c.age |   .9986372   .0000981   -13.89   0.000     .9984449    .9988294
             |
      1.male |   .7964702   .0489286    -3.70   0.000     .7061203    .8983807
   1.insured |   .9599968   .0876252    -0.45   0.655     .8027404     1.14806
       _cons |   .0140093   .0034685   -17.24   0.000     .0086232    .0227595
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Estimating a logit model with the `or` option displays coefficients as odds ratios (the exponent of the estimated coefficients). The odds ratio for `male` from the model, .7964702,

is close to .80696293 which we calculated with predictive margins. The estimates will not be the same because of differences in how the other covariates are fixed. In the calculations above, we left covariates as observed. These are "adjusted" metrics because we include other covariates in the model. (In epidemiology, unadjusted estimates are called "crude" estimates.)

6.7.3 Predictive margins in unadjusted models

With unadjusted models, as in the example in section 6.1.1, using predictive margins and model-based odds ratios produce the same results:

```
logit highchol i.male, nolog or
Logistic regression                             Number of obs     =      5,214
                                                LR chi2(1)        =      12.51
                                                Prob > chi2       =     0.0004
Log likelihood = -3267.8863                     Pseudo R2         =     0.0019
------------------------------------------------------------------------------
    highchol | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.male |   .8104241   .0482285    -3.53   0.000     .7212024    .9106836
       _cons |   .5235662   .0212832   -15.92   0.000     .4834704    .5669872
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.

quietly margins male, post

* marginal effects
display e(b)[1,2] - e(b)[1,1]
-.04573921
* relative risks
display e(b)[1,2] / e(b)[1,1]
.8668999
* odds ratios
display (e(b)[1,2]/(1-e(b)[1,2]) )/ (e(b)[1,1]/(1-e(b)[1,1]))
.81042411
```

6.7.4 Connection between relative risks and odds ratios

After exploring some examples, it should be apparent by now that relative risks are not the same as odds ratios. But in some circumstances they yield similar estimates. Suppose P_t is the probability of an outcome in the treatment group and P_c is the probability in the control group. The odds ratio is:

$$OR = \frac{\frac{P_t}{1-P_t}}{\frac{P_c}{1-P_c}} = \frac{P_t}{Pc} \times \frac{1-P_c}{1-Pt} \tag{6.12}$$

If $\frac{1-P_c}{1-Pt} = 1$, then the odds-ratio is the same as the relative risk $\frac{P_t}{Pc}$. Trivially, this is the case when $P_c \approx P_t$, but also when both are close to zero. This is why relative risks and odds ratios are close when an outcome is a **rare event**. However, odds ratios become significantly larger than relative risks when the probability of the outcome (or **baseline**

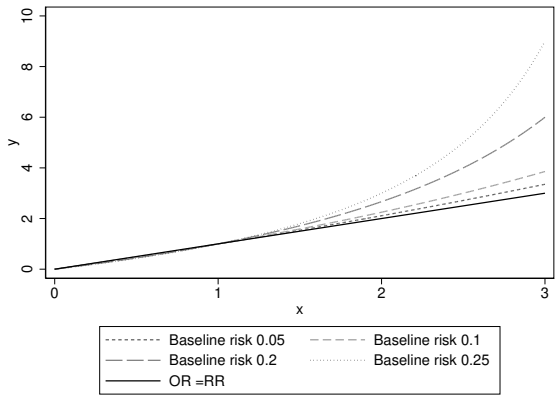

Fig. 6.4 Relationship between odds ratio (y-axis) and relative risk (x-axis)

risk) is larger and $P_c \neq P_t$, which amplifies the problem of interpreting odds ratios as relative probabilities.

The baseline risk in epidemiology is the probability of the outcome at baseline or usually the control group. Rearranging Equation 6.12 in terms of the baseline probability yields, p_0: $OR = \frac{RR - RR * p_0}{1 - RR * p_0}$, following Grant (2014). Figure 6.4 depicts the relationship between odds ratios and relative risk at different baseline probabilities. The odds ratio is plotted on the y-axis and relative risk on the x-axis. The solid black line is the diagonal, in which odds ratios and relative risks are the same. As the baseline risk increases–that is, the outcome is less "rare"– the odds-ratio and the relative risk diverge from the solid diagonal line. The more common the event, the larger the difference. With a baseline risk of 0.25, a relative risk of 3 translates to an odds ratio greater than 8. Thus, it would be very misleading to interpret an odds ratio as a relative risk (see Problem 6.2).

6.8 Probit models

In probit models, marginal effects can be calculated using the same algorithm that we introduced in Section 6.3. The only difference is that predictions are calculated using a different cumulative density function (CDF). As we saw in Chapter 4, the estimation scale in probit models are z scores that need to be converted into a probability. The z scores are given by the **index function** $z = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$. Therefore, predictions in probit models are given by:

$$\Pr(\hat{y}_i = 1 | \boldsymbol{x}_i) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}), \tag{6.13}$$

where $\Phi(\cdot)$ is the **standard normal CDF**. As with logit, predictions can be calculated directly using Stata's `predict` post-estimation command or manually using the standard normal CDF function in Stata, `normal(z)`:

```
quietly probit highchol age i.male i.insured
gen double phat1_norm = normal(_b[_cons] + _b[age]*age
+ _b[1.male]*male + _b[1.insured]*insured)
predict double phat1

sum phat1_norm phat1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  phat1_norm |      5,214    .3212466    .0743214   .1840628   .5044238
       phat1 |      5,214    .3212466    .0743214   .1840628   .5044238
```

Analytically, for a general probit model, the marginal effect for a continuous covariate x_1 is:

$$\frac{\partial \Pr(y_i = 1|\boldsymbol{x})}{\partial x_1} = \beta_1 \phi(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}) \tag{6.14}$$

In Equation 6.14, $\phi(z)$ is the **standard normal probability density function** (PDF), which is the first derivative of the standard normal CDF: $\phi(z)$=$\Phi'(z)$. The direction of the effect is given by β_1. As with the logit model, the marginal effect of x_1 depends on the value of all other covariates in the model even if no explicit interactions were included.

Although at first glance Equation 6.14 does not resemble the marginal effect for the logit model (Equation 6.4), they are in fact related. In Equation 6.4, the term multiplying β_1 is the **standard logistic** PDF. The connection between logit and probit models is clearer in the latent variable approach to derive these models (Chapter 4).

We can write either a probit or logit model as $\Pr(Y = 1|\boldsymbol{x}) = F(\boldsymbol{x'\beta})$ using the general notation, $F(\cdot)$ to denote a CDF (see Notation Box 4.1 for a review of notation). The marginal effect for either model applying the **chain rule** is:

$$\frac{\partial \Pr(y = 1|\boldsymbol{x})}{\partial x_j} = \frac{dF(\boldsymbol{x'\beta})}{d(\boldsymbol{x'\beta})} \frac{\partial(\boldsymbol{x'\beta})}{\partial x_j} = f(\boldsymbol{x'\beta}) \times \beta_j, \tag{6.15}$$

where $f(\cdot)$ is the PDF, which is the first derivative of the CDF. This highlights the fact that predicted probabilities and marginal effects are both based on PDFs – probit using $\phi(\cdot)$ and logit using $\lambda(\cdot)$, the logistic PDF.

If the index function $z = \boldsymbol{x'\beta}$ also includes interaction terms for $\boldsymbol{x}$, for example $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2$; then Equation 6.15 tells us that the marginal effect of x_1 is:

$$\frac{\partial \Pr(y = 1|\boldsymbol{x})}{\partial x_1} = f(\boldsymbol{x'\beta}) \times (\beta_1 + \beta_3 x_2) = \phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2)(\beta_1 + \beta_3 x_2) \tag{6.16}$$

Equation 6.16 highlights an important fact about interaction terms in probit (and logit) models models. The sign of the interaction term in the estimation scale (i.e. the sign of β_3) might not provide information about the direction of effects in the probability scale. The sign in Equation 6.16 is determined by the sign of the expression $(\beta_1 + \beta_3 x_2)$, which depends on all of the terms in the expression, and it could potentially change sign at certain values of x_2. We discuss how to interpret interactions in the probability scale in Section 6.10.

6.8.1 Predicted margins and odds ratios for probit models

Returning to predictive margins, it should be apparent that we can also express probit estimates as odds ratios, relative risks, and risk differences because they can be computed from predictions. As with the logit model, we just use the saved predictive margins to obtain each metric:

```
quietly probit highchol c.age##c.age i.male i.insured
quietly margins male, post

* marginal effects
display e(b)[1,2] - e(b)[1,1]
-.04423684
* relative risks
display e(b)[1,2] / e(b)[1,1]
.87103902
* odds ratios
display (e(b)[1,2]/(1-e(b)[1,2]) )/ (e(b)[1,1]/(1-e(b)[1,1]))
.81608842
```

These are slightly different than the estimates of Section 6.7.2, but note that without covariates, we get the same results as we did with the logit model:

```
quietly probit highchol i.male, nolog
quietly margins male, post

* marginal effects
display e(b)[1,2] - e(b)[1,1]
-.0457392
* relative risks
display e(b)[1,2] / e(b)[1,1]
.86689992
* odds ratios
display (e(b)[1,2]/(1-e(b)[1,2]) )/ (e(b)[1,1]/(1-e(b)[1,1]))
.81042414
```

This example shows that it is indeed possible to obtain odds ratios from probit models. We computed them here for didactic purposes because there is still a widespread misunderstanding regarding the differences between logit and probit models. In the probability scale, probit and logit models are practically the same. One subtle advantage of probit models is that probit models force us to interpret models in the probability scale, which is the scale of interest. On the other hand, we can do the same with logit models. Whether one chooses a normal or logistic CDF to estimate a dichotomous model is about preferences and custom; the difference is usually very small as long as models are interpreted in the probability scale. In economics journals, a logit model is less common than a probit model (but if used, it is commonly interpreted on the probability scale using marginal effects). In medical journals both are used, although logit models are commonly interpreted on the odds scale.

6.9 Marginal effects for other regression models

Although in this chapter we focused on logit and probit models, it is straightforward to apply the definition of marginal effects to other models using the same algorithm described in Section 6.3. In most situations, if predictions can be obtained from a statistical model, then it is possible to calculate marginal effects.

Below, we show how to interpret parameters using the Poisson model introduced in Chapter 4. Poisson and negative binomial models are used to model counts and rates (counts with a denominator). The scale of estimation in the Poisson model is the log of the counts, which can be transformed into relative rates. However, we are often interested in differences of average counts, not relative rates, because differences in average counts give us a better sense of magnitudes.

In this example, we seek to understand if the number of prescription medications taken are associated with having health insurance. We use a sample from the NHANES restricted to those who have high cholesterol and are younger than 65 since most people over age 65 are insured because of Medicare program. (We return to this dataset in Chapter 7, to correct for **overdispertion**, which refers to the outcome having a larger variance than what is expected if the outcome followed a Poisson distribution.)

Below are some basic descriptive statistics:

```
use "nhanes_hyperlipidemia_rx",clear
keep if age < 65
lowess ndrugs insured, jitter(3)

tabstat ndrugs, by(insured) stats(N mean median sd min max)
Summary for variables: ndrugs
     by categories of: insured (1 if any insurance)

 insured |         N      mean       p50        sd       min       max
---------+--------------------------------------------------------------
       0 |       601  .7271215         0  1.588106         0        16
       1 |      2417  1.586678         1  2.339738         0        20
---------+--------------------------------------------------------------
   Total |      3018  1.415507         0  2.236806         0        20
------------------------------------------------------------------------
display 1.586678/.7271215
2.182136
display 1.586678-.7271215
.8595565
```

Figure 6.5 depicts what we expected: those with health insurance use more prescription medications on average than those without health insurance. We could use difference metrics to describe this finding: on average people with health insurance use 2.18 times more prescription medications, or about 0.86 more medications on average (see calculations above). Estimation of a Poisson model with only insurance status yields:

```
poisson ndrugs i.insured, irr nolog

Poisson regression                              Number of obs     =      3,018
```

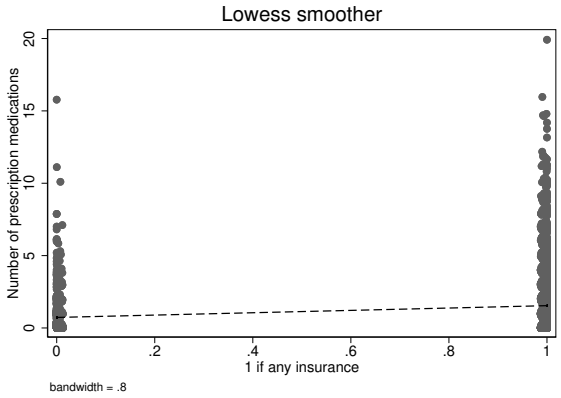

Fig. 6.5 Average prescription medications by health insurance status

```
                                                LR chi2(1)        =     293.35
                                                Prob > chi2       =     0.0000
Log likelihood = -6141.2386                     Pseudo R2         =     0.0233

------------------------------------------------------------------------------
      ndrugs |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.insured |   2.182136   .1101727    15.46   0.000     1.976541    2.409115
       _cons |   .7271215   .0347829    -6.66   0.000     .6620465    .7985929
------------------------------------------------------------------------------
Note: _cons estimates baseline incidence rate.
```

where the option `irr` returns the exponent of the estimated coefficient ("irr" stands for **incidence-rate ratios**) which is the same as the relative rate we computed manually. The effect in terms of average differences can be obtained using the -margins- command with the `dydx()` option:

```
quietly poisson ndrugs i.insured
margins, dydx(insured)

Conditional marginal effects                    Number of obs     =      3,018
Model VCE    : OIM

Expression   : Predicted number of events, predict()
dy/dx w.r.t. : 1.insured

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.insured |   .8595562   .0432009    19.90   0.000      .774884    .9442285
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Because we do not control for other covariates, the effect of insurance expressed in terms of average differences matches our manual calculation.

To complete this discussion, we show more formally how to interpret coefficients from Poisson and negative binomial models with a more general but simple model:

$$log(E[y_i|\boldsymbol{x_i}]) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 D_i, \tag{6.17}$$

where x_1 is a continuous variable and D is a dummy variable. Equation 6.17 matches the estimation scale, so it is our preferred way of writing a Poisson or negative binomial model.

It is straightforward to interpret parameters as relative risks or relative expected counts (incidence-rate ratios) by taking the exponent of model parameters. To see why, we write down Equation 6.17 when $D = 1$, which we denote as $log(E[y1])$, and when $D = 0$, denoted as $log(E[y0])$. We then take the difference and simplify:

$$log(E[y1]) - log(E[y0]) = log\left(\frac{E[y1]}{E[y0]}\right) = \alpha_2 \tag{6.18}$$

After taking the exponent on both sides, we have $E[y1]/E[y0] = e^{\alpha_2}$. Therefore, the exponent of the coefficient can be interpreted as a relative expected count (or a rate if the model includes an offset) holding other covariates constant. You can show that e^{β_1} is the relative expected count when the continuous variable x_1 increases by *one unit* (Problem 6.9; recall that we are working on the log scale).

However, we are often interested in understanding models implications in average counts or rates because these scales match research questions. For a continuous variable, we would could compute the partial derivative $\partial E[y|\boldsymbol{x}]/\partial x_1$. To do this analytically, we could use an **implicit derivative** (see section **??** of Appendix) or by solving for $E[y_i]$ in Equation 6.17 before calculating the derivative. Either way, we obtain:

$$\frac{\partial E[y|\boldsymbol{x}]}{\partial x_1} = \alpha_1 e^{(\alpha_0 + \alpha_1 x_1 + \alpha_2 D)} \tag{6.19}$$

Thus, as with logit and probit models, Equation 6.19 shows that in the count scale the effect of one variable depends on the value of the other variables even if interaction terms are not included in the log scale. When calculating marginal effects using Poisson models, we obtain the numerical (average) derivative corresponding to the analytical derivative of expression 6.19.

6.10 Interaction terms in logit models

In this section, we consider models with explicit interactions in both the odds scale and the probability scale using marginal effects. The notation can get cumbersome, but interpreting interactions in different scales is extremely important in applied work, and marginal effects greatly simplifies the interpretation of interactions.

6.10.1 Interaction terms on the odds scale

In logit models, an interaction term is as a *ratio of odds ratios* in the odds scale. The challenge of interpreting interaction terms in logit models is easier to understand using an example with two indicator variables. Consider the following model, omitting subscripts to simplify the notation:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 HighSchool + \beta_2 male + \beta_3 HighSchool * male \tag{6.20}$$

The variable `HighSchool` equals 1 if a person has at least a high-school degree (we call these individuals "educated" to now simplify the language). The variable `male` equals 1 if the person is male.

On the log-odds scale, we can interpret this model as usual. The effect of having a high-school degree or more on the log-odds of an event depends on sex (and vice versa). The model for males is: $log(\frac{p_m}{1-p_m}) = \beta_0 + \beta_1 HighSchool + \beta_2 + \beta_3 HighSchool$. The model for females is: $log(\frac{p_f}{1-p_f}) = \beta_0 + \beta_1 HighSchool$. The difference between males and females is: $log(\frac{p_m}{1-p_m}) - log(\frac{p_f}{1-p_f}) = \beta_2 + \beta_3 HighSchool$, which shows that the the effect of sex on the log-odds of the outcome depends on the level of education. We can rewrite this model using odds applying properties of logarithms, in particular $log(a) - log(b) = log(a/b)$:

$$\frac{\frac{P_m}{1-P_m}}{\frac{P_f}{1-P_f}} = e^{\beta_2+\beta_3 HighSchool}. \tag{6.21}$$

On the log-odds scale, the interaction is analogous to a **difference of differences** estimate of the impact of education on males versus females, as in linear/OLS. To see this, we can work out the differences:

1. Difference males - females: $log(\frac{p_m}{1-p_m}) - log(\frac{p_f}{1-p_f}) = \beta_2 + \beta_3 HighSchool$
2. Difference male - female for educated: $log(\frac{p_{me}}{1-p_{me}}) - log(\frac{p_{fe}}{1-p_{fe}}) = \beta_2 + \beta_3$
3. Difference male - female for uneducated: $log(\frac{p_{mu}}{1-p_{mu}}) - log(\frac{p_{fu}}{1-p_{fu}}) = \beta_2$
4. The difference of differences (2)-(3) is: $log(\frac{p_{me}}{1-p_{me}}) - log(\frac{p_{fe}}{1-p_{fe}}) - [log(\frac{p_{mu}}{1-p_{mu}}) - log(\frac{p_{fu}}{1-p_{fu}})] = \beta_3$

We can now turn the difference of differences (4) expressed in the log-odds scale into odds. We apply properties of logarithms again as before (twice) to obtain:

$$\frac{\frac{P_{me}}{1-P_{me}}}{\frac{P_{fe}}{1-P_{fe}}} \Big/ \frac{\frac{P_{mu}}{1-P_{mu}}}{\frac{P_{fu}}{1-P_{fu}}} = e^{\beta_3} \tag{6.22}$$

Equation 6.22 tells us that in the odds scale an interaction is a ratio of odds ratios. If greater than 1, the odds of outcome for the educated are higher than for the uneducated, but it would be hard to make any sense of magnitudes. In published research, authors prefer to stratify models by sex or education rather than attempting to interpret a ratio of odds ratios.

With a continuous variable, the logic is the same, but one changes the continuous variable by one unit (see Chapter 4, Section 4.2).

6.10.2 Computing marginal effects with interaction terms

Using marginal or incremental effects, interactions are much easier to understand and to communicate. Returning to our example of high cholesterol, recall that there are biological reasons that suggest that the effect of age on high cholesterol is different for males and females. This can be modeled by including an interaction between age and sex. To simplify the exposition, we drop the insurance variable, which was not statistically significant, and do not add quadratic term for age (Problem 6.7).

We estimate the following model:

$$log\left(\frac{highchol_i}{1-highchol_i}\right) = \beta_0 + \beta_1 age_i + \beta_2 male + \beta_3 age_i \times male_i \tag{6.23}$$

The model in Equation 6.23 is not easy to interpret in the odds scale as we discussed in the previous section. One option would be to stratify by sex to simplify the interpretation (Problem 6.10). However, interpreting models with interactions using marginal effects is much easier. The marginal effects of age for males and females from this model are:

```
quietly logit highchol c.age##i.male
margins, dydx(age) at(male=(0 1))

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
1._at        : male            =           0
2._at        : male            =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0060482   .0004241    14.26   0.000      .005217    .0068794
          2  |   .0007907   .0004604     1.72   0.086    -.0001118    .0016931
------------------------------------------------------------------------------
```

A small increase in age increases the probability of high cholesterol by 0.08 percentage points for males, while age has a larger effect for females, 0.60 percentage points. As always, make sure you carefully read the output. In particular, pay attention to how Stata labels the output. The marginal effect of age for males is labeled as 2 (indicated by the line that starts with `2._at`).

As we noted before, Equation 6.4 –the marginal effect of a continuous variable in the probability scale– shows that the marginal effect depends on the other covariates in the model even if we do not explicitly include interactions. We emphasize this point again here. We estimate marginal effects in a model *without* interactions:

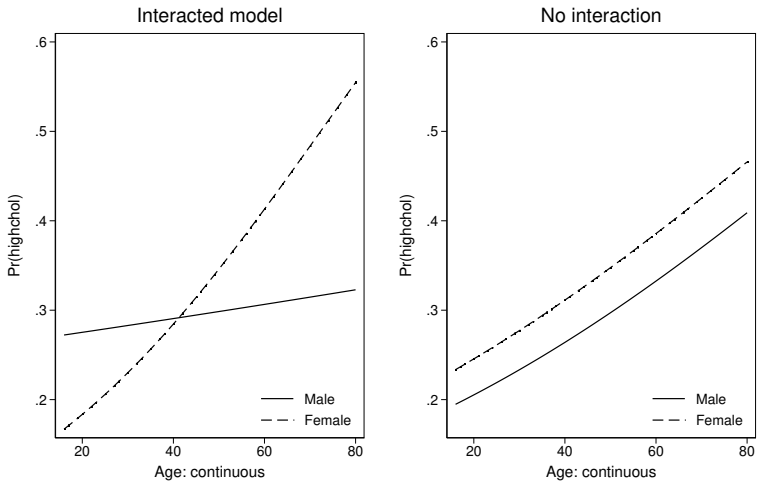

Fig. 6.6 Predicted probability of high cholesterol by sex in models with and without interactions

```
quietly logit highchol c.age i.male
margins, dydx(age) at(male=(0 1)) vsquish

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age
1._at        : male            =           0
2._at        : male            =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0036297   .0003333    10.89   0.000     .0029764    .0042829
          2  |   .0033527   .0003099    10.82   0.000     .0027454    .0039601
------------------------------------------------------------------------------
```

The output above shows two different effects with no interactions, although the difference is not nearly as large as before. One way to think about interactions in logit or probit models is that interactions add another layer of nonlinearity in the probability scale (see Equation 6.16 of the probit model).

Graphically, with interactions, we estimate two separate curves that could cross each other. However, in nonlinear models, even without interactions, the two curves can diverge and can have different average slopes in the probability scale, which is reflected in the marginal effects. Figure 6.6 depicts the difference.

6.11 “Missing” interaction terms in the -margins- output

In the code and output below, we estimate a model with interactions and then request the marginal effects for all the variables in the model. We obtain one marginal effect and one incremental effect, but we estimated a model with three coefficients (excluding the intercept), not two. Why is the marginal effect of the interaction not included? The reason is that Stata is using the algorithms we discussed in Section 6.3, changing the values of age by a small amount and the value of male to 0 and then 1 before making predictions, and it does so in *all terms* in the model that include age or the male variable, including the interaction tern. In other words, the effect of age is an average over the values of sex, and vice versa. Showing effects over the value of the other covariate could be useful to summarize effects, but if we think that interactions are meaningful, we want to evaluate them at specific values of the other covariate using the `at()` option. It is better to always use the factor variable syntax when estimating models that will be interpreted using the -margins- command because then Stata understands that a variable is associated with more than one coefficient in the model:

```
quietly logit highchol c.age##i.male
margins, dydx(*)

Average marginal effects                        Number of obs     =      5,214
Model VCE    : OIM

Expression   : Pr(highchol), predict()
dy/dx w.r.t. : age 1.male
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0034858   .0003117    11.18   0.000     .0028748    .0040967
      1.male |  -.0491543   .0127204    -3.86   0.000    -.0740857   -.0242228
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

If you program interactions manually, you will obtain an uninterpretable marginal effect for the interaction. For example:

```
* Uninterpretable marginal effect for the interaction term
generate age_male = age*male
quietly logit highchol age male age_male
margins, dydx(*)
```

6.12 Standard errors

We have shown how “what if” predictions can be used to calculate numerical derivatives and incremental effects. We can also estimate the standard errors (variance) of the predic-

tions and, by extension, standard errors of marginal and incremental effects (their differences). Predictions are a function of estimated coefficients and the data. In this section, we show two methods to calculate standard errors of predictions.

6.12.1 Direct method for linear/OLS models

Consider first a simple example using a linear model:

$$birthweight_i = \beta_0 + \beta_1 smoke_i + \beta_2 momage_i + \beta_3 smoke \times momage_i + \epsilon_i$$

The outcome is birth weight and the covariates are an indicator for smoking (`smoke`), mother's age (`momage`), and the interaction between the two. We can again use derivatives to understand the implication of the model. What is the effect of mother's age on birth weight? What is the effect of smoking on average birth weight? They are, respectively:

$$\frac{\partial E[birthweight|\boldsymbol{x}]}{\partial momage} = \hat{\beta}_2 + \hat{\beta}_3 smoke \tag{6.24}$$

$$\frac{\Delta E[birthweight|\boldsymbol{x}]}{\Delta smoke} = \hat{\beta}_1 + \hat{\beta}_3 momage \tag{6.25}$$

Each of these expressions is a function of parameters and data (the "hat" is added to the coefficients to emphasize that we are working with estimated coefficients). We need both, coefficient estimates and data on `smoke` and `momage`, to calculate marginal and incremental effects.

To calculate the variance of the effects above, we just need to use basic properties of variances: $var(X + Y) = var(X) + var(Y) + 2cov(X, Y)$, $var(aX) = a^2 var(x)$, and $cov(aX, bY) = ab * cov(X, Y)$. If we apply these properties to the marginal and incremental effects above, we obtain:

$$var\left(\frac{\partial E[bw|\boldsymbol{x}]}{\partial mage}\right) = var(\hat{\beta}_2) + smoke^2 var(\hat{\beta}_3) + 2 * smoke * cov(\hat{\beta}_2, \hat{\beta}_3) \tag{6.26}$$

$$var\left(\frac{\Delta E[bw|\boldsymbol{x}]}{\Delta smoke}\right) = var(\hat{\beta}_1) + mage^2 var(\hat{\beta}_3) + 2 * mage * cov(\hat{\beta}_1, \hat{\beta}_3) \tag{6.27}$$

Therefore, the variance of the marginal effect of age for non-smokers is $var(\hat{\beta}_2)$ and the variance for those who smoked is $var(\hat{\beta}_2) + var(\hat{\beta}_3) + 2 * cov(\hat{\beta}_2, \hat{\beta}_3)$.

These expressions provide a simple way of calculating standard errors of marginal and incremental effects. We first estimate the model and then use the estimated variance-covariance matrix, which is stored by Stata in matrix `e(V)`. The square root of the diagonal elements of the variance-covariance matrix are the standard errors for coefficients reported in the regression output. Since the formulas above involve covariances, we also need some of the off-diagonal elements.

The code below estimates a model and displays the variance-covariance matrix we used

for the calculations. We generate the interaction manually so the matrix $e(V)$ is more compact than the matrix using factor variable syntax.

```
use https://www.stata-press.com/data/r16/cattaneo2, clear
gen sm_age = mbsmoke * mage

qui reg bweight mbsmoke mage sm_age
. matrix list e(V)
symmetric e(V)[4,4]
            mbsmoke        mage      sm_age       _cons
mbsmoke   10733.267
   mage   71.343306   2.6610229
 sm_age  -403.72658  -2.6610229   15.868305
  _cons  -1997.5365  -71.343306   71.343306   1997.5365
```

Using the stored matrix `e(V)` we can then calculate standard errors (Equations 6.26 and 6.27):

```
* Standard error for marginal effect if mbsmoke = 0
display sqrt(e(V)[2,2])
1.6312642
* Standard error for marginal effect if mbsmoke = 1
display sqrt(e(V)[2,2] + e(V)[3,3] + 2*e(V)[3,2])
3.6341825
* Standard error for incremental effect for mage = 25
display sqrt(e(V)[1,1] + (25^2)*e(V)[3,3] + 2*25*e(V)[3,1])
21.555255
```

We can replicate the calculations above using the margins command:

```
quietly reg bweight i.mbsmoke##c.mage
margins, dydx(mage) at(mbsmoke=(0 1)) vsquish

Average marginal effects                        Number of obs     =      4,642
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : mage
1._at        : mbsmoke         =           0
2._at        : mbsmoke         =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mage         |
         _at |
          1  |   11.36258   1.631264     6.97   0.000     8.164523    14.56063
          2  |  -3.950895   3.634182    -1.09   0.277    -11.07562    3.173831
------------------------------------------------------------------------------

margins, dydx(mbsmoke) at(mage=(25)) vsquish
Conditional marginal effects                    Number of obs     =      4,642
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : 1.mbsmoke
at           : mage            =          25
```

```
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |
     smoker  |  -254.0217   21.55526   -11.78   0.000    -296.2802   -211.7631
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

In the margin commands, standard errors are listed in the second column, under "Delta-Method Std. Err." Stata does not use Equations 6.26 and 6.27, but instead uses the **delta method**. In linear models, they are equivalent. However, the delta method is needed when estimating the standard errors of marginal effects in nonlinear models as we discuss in the next section.

6.12.2 Delta method

In this section, we show a more general way to compute standard errors for marginal affects: the **delta method**, but the intuition is the same as in the previous section. The delta method is necessary with nonlinear models because we do not have linear expressions like Equations 6.26 and 6.27 for the variance of marginal and incremental effects so we cannot use the direct method. For the logit model, we would need to calculate the variance of an expression like Equation 6.4. For the probit model, we would need the variance of an expression like Equation 6.14 or 6.16. These expressions are both nonlinear with respect to model coefficients. Fortunately, it is possible to "linearize" these expressions with a first-order **Taylor series expansion** to yield a linearized function and calculate the variance as we did in the previous section.

Consider for example the nonlinear function $f(x) = e^x$, in which x is a random variable. This function can be approximated by a tangent line at point $x = a$ using a first-order Taylor series expansion:

$$f(x = a) \approx f(a) + f'(a)(x - a) = e^a + e^a(x - a) = e^a + e^a x - 2e^a \tag{6.28}$$

It might look like we do not gain much, but in fact we do. The Taylor approximation in Equation 6.28 is linear in x (e^a is just a number), so the variance of the approximation at $x = a$ is $var(e^a + e^a x - 2e^a) = e^{2a} var(x)$. Remember that the variance of a number is zero, so $var(e^a) = 0$ and $var(bx) = b^2 var(x)$. In essence, the Taylor series expansion derives the expression for a line that is tangent to the nonlinear function $y = e^x$ at one specific point $x = a$ (of course, we can change the value of a).

The definition of the derivative is related to the Taylor series expansion, but there is a key difference. The derivative approximates the *slope* of a curve with a tangent line at some point. The first-order Taylor series approximates the *function itself* at some point. (But take the derivative with respected to x in Equation 6.28 and compare it to the derivative of $f(x) = e^x$ evaluated at $x = a$.)

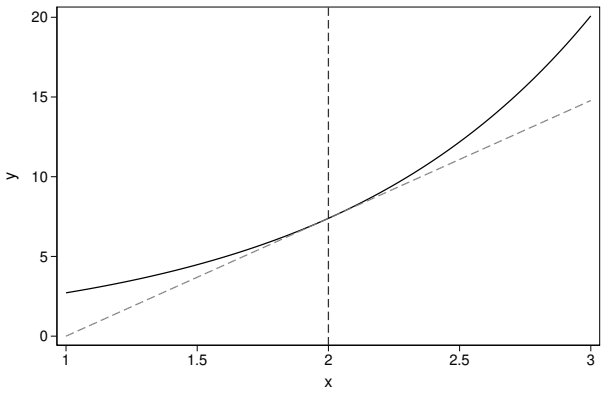

Fig. 6.7 First-order Taylor series approximation of $f(x) = e^x$ at $x = 2$

Figure 6.7 depicts the nonlinear function and and its tangent line at $a = 2$. The straight line is Equation 6.28:

```
twoway function y = exp(x), range(1 3) xline(2, lpattern(dash)) || ///
       function y = exp(2) + exp(2)*(x-2), range(1 3) ///
           color(gray) legend(off)
```

For marginal or incremental effects and predictive margins, we are interested in the variance of a function of random *coefficients* $G(\hat{\boldsymbol{\beta}})$; therefore, we need to extend the first-order Taylor approximation concept to vector functions. We use the variance-covariance matrix along with the vector or matrix of partial derivatives to approximate the variance:

$$var[G(\hat{\boldsymbol{\beta}})] \approx G'(\hat{\boldsymbol{\beta}})cov(\hat{\boldsymbol{\beta}})G'(\hat{\boldsymbol{\beta}})^T \tag{6.29}$$

$G'(\hat{\boldsymbol{\beta}})$ is the matrix of partial derivatives of $G(\hat{\boldsymbol{\beta}})$; that is, the **Jacobian** (or the **gradient** vector), which in this case is the partial derivatives of the predictions with respect to coefficients evaluated at $\hat{\boldsymbol{\beta}}$. $var[G(\hat{\boldsymbol{\beta}})]$ is conditional on the data and the superscript T indicates a matrix transpose.

In the code below, we reproduce Stata's calculation of standard errors for predictive margins using the Jacobian and the variance-covariance matrix. The Jacobian is stored in the matrix `r(Jacobian)` as part of the -margins- command stored results. The key calculation is the computation `Vrep=J*e(V)*J'`, which is Equation 6.29. The option `nofvlabel` displays factor-variable values rather than labels.

```
quietly logit lbweight i.mbsmoke
margins mbsmoke, nofvlabel

Adjusted predictions                            Number of obs     =      4,642
Model VCE    : OIM

Expression   : Pr(lbweight), predict()

------------------------------------------------------------------------------
```

```
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     mbsmoke |
           0 |   .0489677   .0035109    13.95   0.000     .0420864     .055849
           1 |   .1099537   .0106428    10.33   0.000     .0890943    .1308131
------------------------------------------------------------------------------

* Save Jacobian from margins in J
matrix J = r(Jacobian)
* Delta method variance
matrix Vrep = J*e(V)*J'
* Match margins output SE
display sqrt(Vrep[1,1])
.00351092
display sqrt(Vrep[2,2])
.01064276
```

The relevant concept to understand is that the standard errors of marginal and incremental effects measure the variability of the predictions we use to interpret model implications. In simple linear models, because of additive and separable effects, the standard errors of marginal effects are the same as the standard errors of the corresponding coefficient. However, they are usually different in nonlinear models like logit or probit.

The part that we left unexplained is how Stata forms the Jacobian. In this example, the rows of the Jacobian matrix correspond to the two predictive margins and the columns correspond to the partial derivatives with respect to the coefficients $\hat{\boldsymbol{\beta}}$ of the predictive margins. Therefore, the Jacobian matrix has two rows (two predictive margins) and two columns since there are two coefficients, $\hat{\beta}_1, \hat{\beta}_2$. However, when replicating Stata's calculation of the Jacobian below, keep in mind that Stata's factor variable syntax includes the baseline for `smoked` (column of zeroes), so the Jacobian will be a 2x3 matrix.

To simplify the notation, the estimated model is $log(\frac{lw}{1-lw}) = \beta_0 + \beta_1 s$, where lw is the outcome `lbweight` and s is `mbsmoke`. The two predicted margins are:

$$p_0 = \Pr(lw = 1|s = 0) = \frac{1}{1 + e^{-\hat{\beta}_0}} \tag{6.30}$$

$$p_1 = \Pr(lw = 1|s = 1) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1)}} \tag{6.31}$$

Therefore, the Jacobian is:

$$J = \begin{pmatrix} \frac{\partial p_0}{\partial \beta_1} & \frac{\partial p_0}{\partial \beta_0} \\ \frac{\partial p_1}{\partial \beta_1} & \frac{\partial p_1}{\partial \beta_0} \end{pmatrix} = \begin{pmatrix} 0 & (1 - p_0)p_0 \\ (1 - p_1)p_1 & (1 - p_1)p_1 \end{pmatrix} \tag{6.32}$$

The calculus and algebra needed to derive matrix J are not difficult. Below, we show one of them in steps. Do not forget the **chain rule** and the fact that $1/(1 + e^{-\hat{\beta}_0}) = (1 + e^{-\hat{\beta}_0})^{-1}$:

$$\frac{\partial p_0}{\partial \beta_0} = \frac{e^{-\hat{\beta}_0}}{(1 + e^{-\hat{\beta}_0})^2} = \frac{e^{-\hat{\beta}_0}}{(1 + e^{-\hat{\beta}_0})} \times \frac{1}{(1 + e^{-\hat{\beta}_0})} = (1 - p_0) \times p_0$$

Below is the code to obtain the Jacobian matrix, which we compare to the one saved by

the -margins- command (with an extra column for the baseline category of `smoked`). Note that we use the inverse logit function `invlogit` to calculate predictions. It is the same as using Equations 6.30 and 6.31:

```
quietly logit lbweight i.mbsmoke
* Predictions
scalar p0 = invlogit(_b[0b.mbsmoke] + _b[_cons])
scalar p1 = invlogit(_b[1.mbsmoke] +  _b[_cons])
* Jacobian matrix
matrix J = (0,(1-p0)*p0\(1-p1)*p1,(1-p1)*p1)
. matrix list J
J[2,2]
           c1         c2
r1          0  .04656987
r2  .09786389  .09786389

* Compare to margins
quietly logit lbweight i.mbsmoke
quietly margins mbsmoke
. matrix list r(Jacobian)
r(Jacobian)[2,3]
            lbweight:  lbweight:  lbweight:
                  0b.         1.
             mbsmoke    mbsmoke      _cons
0.mbsmoke          0          0  .04656987
1.mbsmoke          0  .09786389  .09786389
```

To calculate the standard error of the incremental effect of `smoked`, the procedure is the same, except that the incremental effect is the difference between the two predictions, $(p_1 - p_0)$, so we would need to calculate the partial derivatives of the difference with respect to model coefficients. We could also use the delta method to calculate standard errors of other metrics, like relative risks, which is just $\frac{p_1}{p_0}$, and of course, marginal effects. The delta method is a flexible tool that can be used to approximate the variance of functions of random variables at some particular point.

It is not difficult to see why the coefficients' standard errors in a linear model (with linear terms for all covariates) are the same as the delta method's standard errors. When computing marginal effects, the Jacobian will consist of partial derivatives of linear functions with respect to coefficients, so we end up with a vector with elements that are 1 for the coefficient of interest and 0 for the rest, which are then multiplied by the variance-covariance matrix and the transpose of the vector with elements 1 and 0. The delta method formula (Equation 6.29) would "pick up" the variance of the corresponding coefficient from the variance-covariance matrix. In the example below, the multiplication `J*e(v)` "selects" the third row of the variance-covariance matrix. When `J*e(v)` is multiplied by `J'`, the element `e(V)[3,3]` is obtained, which is the just variance of the coefficient of `mage` from the variance-covariance matrix:

```
quietly reg lbweight i.mbsmoke mage
* Variance of beta_age
. display e(V)[3,3]
3.878e-07
quietly margins, dydx(mage)
matrix J = r(Jacobian)
```

```
matrix list J
J[1,4]
           0b.       1.
      mbsmoke  mbsmoke     mage    _cons
mage        0        0        1        0

* Delta method variance same as coefficient variance
matrix Vrepl = J*e(V)*J'
matrix list Vrepl
symmetric Vrepl[1,1]
           mage
mage  3.878e-07
```

6.13 Interaction terms in nonlinear models: a special case

Interpreting interactions on the probability scale raises interesting issues that highlight the difference between the estimation scale and the scale of interest in nonlinear models. We focus on the probit model and use the same data used in Karaca-Mandic et al. (2012), although we analyze the data in a different way. As we will see, these data highlight a case when intuition based on linear regression does not extend to nonlinear models. The issues discussed in this section are in the context of a probit model but apply more generally to the logit model and other nonlinear models, like GLM models with a non-identity link.

Suppose we estimate a probit model with age and sex and their interaction as covariates. The model is:

$$\Pr(y_i = 1|\boldsymbol{x}_i) = \Phi(\beta_1 + \beta_2 age_i + \beta_3 age_i \times female_i)$$

```
webuse margex, clear
generate female = (sex==1)

sum outcome age female
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     outcome |      3,000    .1696667    .3754023          0          1
         age |      3,000      39.799    11.54174         20         60
      female |      3,000    .5006667    .5000829          0          1

probit outcome c.age##i.female, nolog

Probit regression                               Number of obs     =      3,000
                                                LR chi2(3)        =     564.73
                                                Prob > chi2       =     0.0000
Log likelihood = -1083.7069                     Pseudo R2         =     0.2067

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0560815   .0053044    10.57   0.000      .045685     .066478
```

```
    1.female |   .4183178   .3065274     1.36   0.172     -.182465      1.0191
             |
female#c.age |
          1  |    .001315   .0066631     0.20   0.844    -.0117445    .0143745
             |
       _cons |  -3.691749   .2371119   -15.57   0.000     -4.15648   -3.227018
------------------------------------------------------------------------------
```

The output above shows that the interaction term is not statistically significant, with a p-value = 0.844. We could reach the same conclusion with a **likelihood ratio test** comparing a model with the interaction term to a model without the interaction (remember that the Wald test and the likelihood ratio test are **asymptotically equivalent**):

```
quietly probit outcome c.age##i.female, nolog
estimate store m_full
quietly probit outcome c.age i.female, nolog
estimate store m_reduced
lrtest m_full m_reduced

Likelihood-ratio test                                 LR chi2(1)  =      0.04
(Assumption: m_reduced nested in m_full)              Prob > chi2 =    0.8437
```

Both tests suggest the same conclusion: model fit is not improved when the interaction term is added to the model (recall that the null in the likelihood ratio test is that the reduced model is adequate; we do not reject the null). The common modeling approach would be to not include the interaction term in the model, although on some occasions scientific knowledge of the **data-generating process** would lead us to include an interaction regardless of model fit. However, let's use marginal effects to understand model implications in the probability scale *keeping* the interaction term:

```
quietly probit outcome c.age##i.female, nolog
margins, dydx(age) at(female=(0 1)) vsquish

Average marginal effects                        Number of obs     =      3,000
Model VCE    : OIM

Expression   : Pr(outcome), predict()
dy/dx w.r.t. : age
1._at        : female          =           0
2._at        : female          =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |    .009059   .0009752     9.29   0.000     .0071478    .0109703
          2  |    .013659   .0006922    19.73   0.000     .0123023    .0150157
------------------------------------------------------------------------------
```

The output above shows that the marginal effect of age on the *probability* of the outcome is different for males and females. The marginal effect for females is .013659, which is higher than the marginal effect for males of .009059. Notice something else: the confidence intervals do not overlap. This would suggest that the effect of age on the probability of the

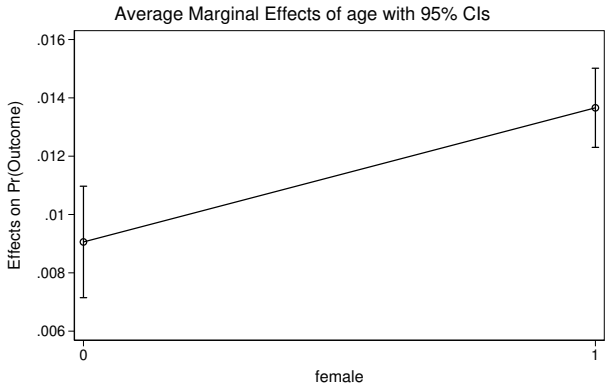

Fig. 6.8 Marginal effects by sex with confidence intervals

outcome is statistically different for males than females. We can see this graphically as well using the `marginsplot` postestimation command (Figure 6.8):

```
quietly probit outcome c.age##i.female, nolog
margins, dydx(age) at(female=(0 1)) post
marginsplot
```

We could also perform a test of the statistical significance of marginal effects (using the delta method to estimate the variance) with the `contrast` option:

```
quietly probit outcome c.age##i.female, nolog
margins, dydx(age) at(female=(0 1)) contrast(at) vsquish

Contrasts of average marginal effects           Number of obs     =      3,000
Model VCE    : OIM

Expression   : Pr(outcome), predict()
dy/dx w.r.t. : age
1._at        : female          =           0
2._at        : female          =           1
-----------------------------------------------
             |         df        chi2     P>chi2
-------------+---------------------------------
age          |
         _at |          1       14.80     0.0001
-----------------------------------------------
```

We are now in a vexing situation. We have two seemingly contradictory answers about interactions in different scales. On the estimation scale, the interaction does not improve model fit but on the probability scale, the variability of the counterfactual prediction indicates that the marginal effect of age is different for males and females. However, there is something else we should do, which will help us resolve the apparent contradiction. The likelihood ratio test (and the asymptotically equivalent Wald test) suggests that the interaction term does *not* improve model fit. So let's explore the implications of a model *without* the interaction term in the probability scale:

```
quietly probit outcome c.age i.female, nolog
margins, dydx(age) at(female=(0 1)) vsquish

Average marginal effects                                    Number of obs = 3,000
Model VCE: OIM
Expression: Pr(outcome), predict()
dy/dx wrt:  age
1._at: female = 0
2._at: female = 1

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |    .009193   .0007025    13.09   0.000     .0078161      .01057
          2  |   .0135865   .0005891    23.06   0.000      .012432    .0147411
------------------------------------------------------------------------------
```

Once again, we find that the marginal effects are different even without the interaction term. We can test the difference again:

```
quietly probit outcome c.age i.female, nolog
margins, dydx(age) at(female=(0 1)) contrast(at) vsquish
Contrasts of average marginal effects                       Number of obs = 3,000
Model VCE: OIM

Expression: Pr(outcome), predict()
dy/dx wrt:  age
1._at: female = 0
2._at: female = 1

------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
age          |
         _at |          1       56.88     0.0000
------------------------------------------------
```

We arrive at identical conclusions as with the model with interactions. On the probability scale, effects are no longer additive or separable, therefore, the effect of age depends on sex *even if the model does not explicitly include an interaction term* (see Equation 6.14).

For more intuition, we plot predicted probabilities for males and females in different scales with and without interactions (Figure 6.9):

```
quietly probit outcome c.age##i.female, nolog
predict phat
predict zhat, xb
quietly probit outcome c.age i.female, nolog
predict phat_ni
predict zhat_ni, xb
* See the online supplemental material for the code to replicate graph
```

Figure 6.9 shows that models with and without an interaction term are similar when compared on the same scale (z scores and probability). Clearly, the effects between males and

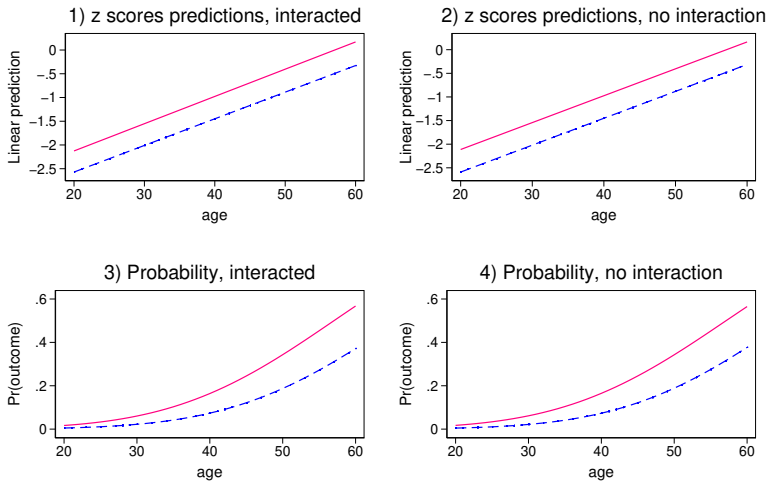

Fig. 6.9 Predicted outcome in different scales by sex; dashed line is males

females are different on the probability scale with or without an interaction term. Adding the interaction term does not appear to change predictions, which agrees with tests of model goodness of fit and the exploration of marginal effects. On the probability scale, as age increases, the probability of the outcome is larger for females. Visually, at age 20, the outcome appears to have a similar low probability. At age 60, the probability for females is much higher than for males. We can use marginal effects to calculate effects in the probability scale without interactions at different ages:

```
quietly probit outcome c.age i.female, nolog
margins, dydx(age) at(female=(0 1) age=(20)) vsquish
Conditional marginal effects                    Number of obs     =      3,000
Model VCE    : OIM

Expression   : Pr(outcome), predict()
dy/dx w.r.t. : age
1._at        : age             =          20
               female          =           0
2._at        : age             =          20
               female          =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0007937   .0001552     5.11   0.000     .0004894    .0010979
          2  |   .0024387   .0003679     6.63   0.000     .0017176    .0031597
------------------------------------------------------------------------------

margins, dydx(age) at(female=(0 1) age=(60)) vsquish
Conditional marginal effects                    Number of obs     =      3,000
```

```
Model VCE    : OIM

Expression   : Pr(outcome), predict()
dy/dx w.r.t. : age
1._at        : age             =          60
               female          =           0
2._at        : age             =          60
               female          =           1
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   .0216201   .0016145    13.39   0.000     .0184557    .0247844
          2  |   .0224025   .0011217    19.97   0.000     .0202039    .0246011
------------------------------------------------------------------------------
```

We calculated marginal effects at 20 and 60 separately because the output is easier to read than calculating them together (`margins, dydx(age) at(female=(0 1) age=(20 60))`). To calculate the *predicted probabilities* at different ages, we could use predictive margins with:

```
margins female, at(age=(20 60)) vsquish
```

The seemingly contradictory findings are clarified when we distinguish model fit from exploration of effects in different scales.

To provide further intuition, consider fitting a linear probability model instead. Without an interaction term, the effect of age would be the same for males and females. With an interaction term, we *explicitly* allow different effects of age for sex, thus the predicted probabilities (straight lines if age is entered linearly) could cross each other. It is like estimating two stratified models, one for females and one for males. In the linear probability model, the interaction is statistically significant:

```
quietly {
   regress outcome c.age i.female, robust
   estimate store lpm_noint
   reg outcome c.age##i.female, robust
   estimate store lpm_int
   predict yhat_int
}
estimates table lpm_noint lpm_int, star
----------------------------------------------
    Variable |   lpm_noint        lpm_int
-------------+--------------------------------
         age |  .01135301***    .00752229***
             |
      female |
          1  |  .09640036***   -.20263792***
             |
female#c.age |
          1  |                   .0075271***
             |
       _cons |  -.3304364***   -.19213642***
```

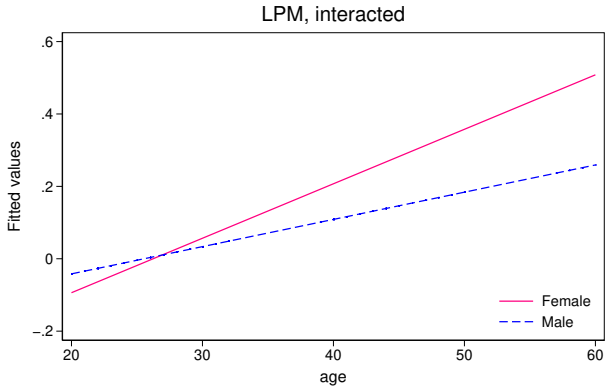

Fig. 6.10 Predicted probability by sex (interacted model)

```
------------------------------------------------
      legend: * p<0.05; ** p<0.01; *** p<0.001
```

Figure 6.10 shows predicted probabilities by sex for the linear probability model. Model fit increases because the evidence suggests that the best fitting model is one that allows the lines to have different slopes. With a probit model, we did not need an explicit interaction term to obtain a well-fitted model because even without an interaction term the effect of one variable depends on the value of other covariates.

6.13.1 Which standard error should we use to test interaction terms?

The question that remains is which standard error should be used to decide if the interaction term is retained in the model. Karaca-Mandic et al. (2012) separated the question of goodness of fit (i.e. Wald test or likelihood ratio test) of the interaction term from the question of testing the marginal effect at different values of the other covariate(s). For the latter, they recommend using a test based on the delta method standard errors (as with the `contrast` option above). However, Greene (2010) provides a different point of view. He suggests building models based on goodness of fit and other traditional tools of modeling to assess model specification. He argues that marginal effects are "neither coefficients nor elements of the specification of the model. They are implications of the specified and estimated model.... [It] seems it would be rare for a model builder to build a structural model by hypothesizing (statistically) about partial effects and/or predictions that would be made by that model." We agree with Greene's interpretation of marginal effects as model implications, although model implications may change depending on the scale.

In Chapter 11, we discuss implications of interactions in difference-in-difference designs with nonlinear models. In difference-in-difference, the coefficient of interest is the interaction term and the test of a key assumption – parallel pre-trends– is also a test of the statistical significance of an interaction term. As the example of Section 6.13 shows,

we could reach different conclusions about interactions if we do not explore the possibility that interactions could be present in the scale of interest even when the interaction term is not statistically significant in the estimation scale. Furthermore, in difference-in-difference designs, the *identification* of causal effects also depends on the scale.

6.14 Additional topics

This section briefly presents some additional topics that are important to keep in mind but not fundamental for using marginal effects in practice.

6.14.1 Numerical precision

Numerical precision is vital to obtain accurate marginal effects. We discussed the choice of h in Section 6.2. The relevance of this choice is easier to understand with an example. Suppose we want to calculate the numerical derivative of $y = e^x$ at $x_0 = 2$. We know the correct answer is $\frac{dy}{dx}\Big|_{x=2} = e^x = e^2$. If we set $h = 10^{-8}$, we can tell that our numerical derivative approximation is correct up to the first 8 digits:

```
clear
set obs 1
generate double y = exp(2)
generate double dydx = (exp(2+1e-8) - exp(2-1e-8))/(2*1e-8)
display %20.0g y
7.38905609893065041
display %20.0g dydx
7.3890560337019906
```

The source of error is the difference in the numerator of `dydx` above. This is why choosing the best "small but large enough" h is important. If h is too small, then the numerator of Equation 6.7 would be essentially zero, which means we cannot obtain the numerical approximation of the derivative. If too large, the approximation is not accurate. The -margins- command uses an iterative procedure, first starting at the initial value of h to find a value of h that maximizes numerical accuracy. Numerical accuracy in this context is defined as a comparison of $f(x_0 + h)$ and $f(x_0)$. Optimal numerical accuracy is achieved if they differ in about half their digits. See (Gould et al., 2006, Chapter 1) for more on this example and Thisted (2017) for an introduction to numerical computation in statistics.

6.14.2 Is it a unit change?

From the discussion in this chapter, it should be clear that marginal effects are effects due to small changes in the value of a covariate, not a unit change, as they are often described. They are unit changes for incremental effects because we code indicator variables as 1 or 0. However, it can be confusing to then read that marginal effects are changes in average outcome when the value of a covariate changes by "one unit" – for example, one additional

year of life. A more precise statement is that marginal effects are *approximations* of the change in the outcome when age changes by 1 year.

Whether there is a difference in the change between 1 unit and the marginal effects *approximation* depends on whether the relationship between x and $E[y|\boldsymbol{x}]$ is linear or not. To develop some intuition with a simple example, consider the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$. We know how a very small change in x changes average outcome: $\frac{dE[y_i|x]}{dx} = (\beta_1 + 2\beta_2 x) \times \Delta x$. We can use the definition of the analytical derivative to derive the previous formula when x changes by a small amount h that tends to zero:

$$\lim_{h \to 0} \frac{\beta_0 + \beta_1(x+h) + \beta_2(x+h)^2 - \beta_0 - \beta_1 x - \beta_2 x}{h} = \beta_1 + 2\beta_2 x + \beta_2 h$$

When $h \to 0$, the term $\beta_2 h$ vanishes. But if the change were a unit change, $h = 1$, then $\beta_2 h = \beta_2$. This gives us some insight: a very small change and a unit change will be the same if $\beta_2 = 0$. The parameter β_2 is the coefficient of x^2, the nonlinear term in the model, so both are going to be the same if the relationship between $E[y|\boldsymbol{x}]$ and x is a straight line, and they will differ when there is more curvature.

To obtain a unit change at a particular point (or on average), simply set $h = 1$ in step 2 of the algorithm of Section 6.3. Standard errors can be calculated using the delta method.

6.14.3 Margins over populations

To produce Figure 6.9, we calculated model predictions and plotted predicted lines by sex (see the Stata code corresponding to this chapter). However, it also possible to use the -marginsplot- postestimation command to quickly plot predictions taking advantage of the `over()` option, which estimates marginal effects or margins at unique values of the variable(s) specified in `over()`. For example, we could reproduce the third panel of Figure 6.9, titled "3) Probability, interacted," with:

```
webuse margex, clear
generate female = (sex==1)

quietly probit outcome c.age##i.female
quietly margins, over(female) at(age=(20(5)60))
marginsplot
```

6.14.4 Nonparametric models

We have presented marginal effects as tools to interpret model *parameters*, but marginal effects can also be used to interpret **nonparametric models**. In a model like $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji} + \epsilon_i$, the parameters $\boldsymbol{\beta}$ are estimated (j+1 parameters). The expected value of the outcome is a function of parameters for given values of covariates: $E[y_i|\boldsymbol{x}_i] = f(\boldsymbol{x}_i, \boldsymbol{\beta})$, which is how we calculate predictions. Above, we made assumptions about the shape of $f(\cdot)$; all variables are additive and linear. In nonparametric models, the expected value is a function of covariates but not parameters, $E[y_i|\boldsymbol{x}_i] = g(\boldsymbol{x}_i)$. Furthermore, the function $g(\cdot)$ is not known and is assumed to be an element of an infinite-dimensional

space of functions, so we do not make assumptions about the shape of $g(\cdot)$. Rather than estimating parameters, $E[y_i|\boldsymbol{x}_i] = g(\boldsymbol{x}_i)$ is estimated conditional on the values of covariates. The estimated function $E[\hat{y}_i|\boldsymbol{x}_i]$ can be used to interpret the nonparametric model, including calculating how changes in the value of covariates change $E[\hat{y}_i|\boldsymbol{x}_i]$.

The example below illustrates the use of -margins- with the -npregress kernel- command, which by default estimates a local linear regression, with bootstrapped standard errors. Marginal and incremental effects are interpreted as usual. An increase in age increases average total serum cholesterol by 0.30 mg/dl, holding sex constant. Total serum cholesterol is 7.45 mg/ld less on average for males, holding age constant.

```
use nhanes_cholesterol,clear
quietly npregress kernel tcresult age i.male
margins, dydx(*) vce(bootstrap)
(running margins on estimation sample)

Bootstrap replications (50)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
Average marginal effects                                Number of obs = 5,214
                                                        Replications  =    50

Expression: Mean function, predict()
dy/dx wrt:  age 1.male
------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Percentile
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .3013848    .043669     6.90   0.000     .2083199    .3836323
      1.male |  -7.452098   .9628852    -7.74   0.000    -9.465876   -5.489199
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

* Compare to parametric counterpart:
regress tcresult age i.male
```

6.14.5 Margins for transformed covariates

Throughout this chapter, we emphasized the use of the factor variable syntax to estimate models in order to obtain correct marginal effects in Stata. The factor syntax is flexible. For example, a model with linear and quadratic terms for age can be estimated with `c.age##c.age` or `c.age c.age#c.age`. A model adding a cubic term (i.e., polynomial of third degree) is estimated with:

```
regress tcresult i.male c.age c.age#c.age c.age#c.age#c.age
```

However, not all transformations can be specified with the factor syntax. For example, a model with log(age) as a covariate needs to be specified with a transformed variable:

```
generate log_age = log(age)
regress tcresult log_age
```

One consequence of "manually" transforming a variable is that `margins, dydx(log_age)` would compute the numerical equivalent of $\frac{dE[tcresult|age]}{dlog(age)}$ rather than $\frac{dE[tcresult|age]}{dage}$, which is

likely of interest. For example, in the model $tcresult_i = \beta_0 + \beta_1 log(age_i) + \epsilon_i$, the derivative is $\frac{dE[tcresult|age]}{dage} = \beta_1 \frac{1}{x}$. The option `expression()` can be used to obtain marginal effects and standard errors with respect to age:

```
regress tcresult log_age
margins, expression(_b[log_age]*(1/age))
Predictive margins                                      Number of obs = 5,214
Model VCE: OLS
Expression: _b[log_age]*(1/age)
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   .4320437    .030629    14.11   0.000      .372012    .4920754
------------------------------------------------------------------------------
```

When age changes by one year, the average serum cholesterol increases by *approximately* 0.43 mg/ld. The above marginal effect is the same as evaluating the analytical derivative at all observed ages and then taking the average:

```
regress tcresult log_age
generate pred1 = _b[log_age]*(1/age)
```

Another alternative is to obtain the derivative numerically, modifying the algorithm in Section 6.3 by adding and subtracting h from age –rather than $log(age)$– and then taking the logarithm to obtain the numerical approximation of $\frac{dE[tcresult|age]}{dage}$ instead of $\frac{dE[tcresult|age]}{dlog(age)}$. The online supplemental material for this chapter presents an example.

6.15 Summary

This chapter presented marginal and incremental effects as tools to interpret regression models – parametric and nonparametric. Marginal effects approximate analytical derivatives and incremental changes with their numerical counterparts, which are computed using what-if model predictions. We emphasized that marginal effects are rates of change, and since marginal effects rely on model predictions, the predictions (i.e., predictive margins) can also be used to calculate different effect metrics, like relative risks with logit and probit models. We also emphasized that most common nonlinear models are estimated in a scale that is often not the scale of interest. This chapter showed how marginal effects and predictive margins simplify model interpretation in different scales. The issue of different scales allowed us to explore the counterintuitive implications of nonlinear models, particularly in models with interaction terms. We will extensively use marginal effects in the rest of the book.

6.16 Further readings and additional material

Stata's reference manual for the -margins- command is the best source for further options and details. The Methods and Formulas section of the Stata manual for -margins- provide a more general definition of marginal effects. However, precise definitions are model dependent. In quantile regression (command -qreg-), for example, marginal effects are model implications on the *median* of the outcome rather than the mean.

We do not cover other features like elasticities and semielasticities, which are additional metrics to understand model implications, but we have additional material on elasticities on our website. The online supplemental material includes examples using -predictnl- and -lincom- to calculate standard errors for marginal effects and a replication of the analysis in Greene (2010) using analytical derivatives to obtain average marginal effects.

For more details on derivatives, Simon and Blume (1994) and Spivak (2006) are excellent sources. Wooldridge (2010) Chapter 15, and Greene (2018), Chapter 17, are good sources on probit and logit models, including derivations of analytical marginal and incremental effects. We have used a combination of their notation so you can follow their presentation.

Problems

6.1 Suppose you estimate the following linear/OLS model: $y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 z_i + \epsilon_i$, where x is a continuous variable and z is an indicator or dummy variable taking only two values, 0 and 1. You estimate the model with `regress y x i.z`. Explain what the following commands would calculate:

a). `margins, dydx(*)`
b). `margins z`
c). `margins, at(x=(0(1)20))`
d). `margins, at(x=20 z=1)`

6.2 In Section 6.7.2, we used a simple numerical example to calculate relative risks, relative differences, and odds ratios. In Section 6.7.4, we showed that effects with odds ratios are much larger than relative risk when the baseline risk is high. Repeat the calculations in Section 6.7.2 assuming that the probability of the outcome in the control group is 0.9 and 0.95 in the treated group. Show that the odds ratio imply a 111 percent increase for the treatment group, while the relative risk implies a modest 5.6 percent increase.

6.3 Explain the difference between Figure 6.2 and Figure 6.3.

6.4 Figure 6.1 suggests that the relationship between age and the probability of high cholesterol is nonlinear. In a model with sex as a covariate, explore different logit models specifications for age. For each model, use marginal effects for age to understand model implications in the probability scale. Decide which specification is

better using Wald tests, likelihood ratio tests, or BIC for non-nested models. Plot predictive margins to better understand each model.

a). Start with a model with age entering linearly in the log-odds scale.
b). Include a quadratic term for age.
c). Include a quadratic and cubic term for age.
d). Estimate a model with $log(age)$ as predictor.
e). Estimate a model with $\sqrt{age}$ as predictor.
f). Estimate a model using indicators for age in quintiles (so a model with four dummy variables).

6.5 Use the nhanes_cholesterol.dta dataset. The variable `tcresult` records total cholesterol in milligrams per decilitre (mg/dL). Estimate the following linear/OLS model: $tcresult_i = \gamma_0 + \gamma_1 age_i + \gamma_2 male_i + u_i$.

a). Use *analytical* partial derivatives and incremental changes to interpret the effect of age and sex on $E[tcresult_i|age_i, male_i]$.
b). Repeat a) using *numerical* marginal and incremental effects. Why are the results the same as in a)?
c). Estimate now the following model: $tcresult_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 (age \times male)_i + \epsilon_i$.
  c1). Interpret the results of: `margins, dydx(age)`
  c2). Interpret the results of: `margins, dydx(age) at(male=(0 1))`
  c3). In c), age enters the model linearly. Show that including a quadratic term for age fits the model better. Repeat c1) and c2). Does the interpretation of the -margins- findings change?
  c4). Plot predictive margins for age (16 to 60 in increments of 5) for the model in c3) (with a quadratic term for age). Use the `over(male)` option to plot predictive margins separately for males and females. Interpret the plot.

6.6 Suppose you have the following logit model: $log(\frac{p_i}{1-p_i}) = \alpha_0 + \alpha_1 x_{1i} + \alpha_1 x_{2i}$. Solve for p and show the expression to predict the probability of the outcome when $x_1 = a$ and $x_2 = b$.

6.7 Use the dataset nhanes_cholesterol.dta for this problem. Use age, male, and insured as covariates.

a). Show that a quadratic term for age is justified.
b). Does the previous answers change when removing insured and sex from the model?
c). Graph *predictions* using your preferred model specification.
d). Graph *marginal effects* for age using your preferred specification.
f). Evaluate marginal effects at different ages and graph them. Describe your findings.
g). Include interactions between age and sex in a model with age modeled with a quadratic term. Graph marginal effects and predictions for males and females in the same graph (separate lines for each sex).

6.8 Problems set using individual derivatives.

6.9 In Section 6.9, we presented the Poisson model $log(E[y_i|\boldsymbol{x_i}]) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 D_i$, where x_1 is a continous variable and D is a dummy variable. We showed that the exponent of α_2 is a relative risk. Show that the exponent of α_1 is a also a relative risk when x_1 changes by one unit. (Hint: write down the model with $(x_1 + 1)$ and the model with x_1 and take the difference).

6.10 Show the interacted model in 6.10.2 is equivalent to models stratified by sex (log odds and odds ratio scales).

6.11 Use the dataset nhanes_cholesterol.dta for this problem. Estimate the following logit model for females only: $logit(highchol_i) = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + \beta_3 insured_i$. We saw that the probability of high cholesterol appears to peak around age 55 (combining males and females). Find the exact peak in the probability scale for females using the estimated model results (that is, parameters). To make it easier, calculate the peak for a female who is uninsured.