Health Econometrics Using Stata

Partha Deb
Hunter College, CUNY and NBER

Edward C. Norton
University of Michigan and NBER

Willard G. Manning
University of Chicago

A Stata Press Publication
StataCorp LLC
College Station, Texas

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

Print ISBN-10: 1-59718-228-1
Print ISBN-13: 978-1-59718-228-7
ePub ISBN-10: 1-59718-229-X
ePub ISBN-13: 978-1-59718-229-4
Mobi ISBN-10: 1-59718-230-3
Mobi ISBN-13: 978-1-59718-230-0

Library of Congress Control Number: 2016960172

Dedication to
Willard G. Manning, Jr. (1946–2014)

Will Manning joined the RAND Corporation in 1975, a few years after completing his PhD at Stanford. He quickly became involved in the RAND Health Insurance Experiment. Will was the lead author of the article that reported the main insurance results in the 1987 *American Economic Review*, one of the most cited and influential articles in health economics. He also published many seminal articles about demand for alcohol and cigarettes, the taxes of sin, and mental healthcare. In 2010, the American Society of Health Economics awarded the Victor R. Fuchs Award to Will for his lifetime contributions to the field of health economics.

But perhaps his strongest influence was on empirical methods central to applied health economics research. With others at RAND, he advocated moving away from tobit and sample-selection models to deal with distributions of dependent variables that had a large mass at zero. The two-part model, in all of its forms, is now the dominant model for healthcare expenditures and use. He also understood the power and limitations of taking logarithms of skewed distributions. And woe to the author who did not deal adequately with heteroskedasticity upon retransformation. Will continued to push the field of health econometrics through the end of his career. He helped develop new methods and advocated the work of others who found better ways of modeling healthcare expenditures and use. His influence on applied health economics is deep and lasting (Konetzka 2015; Mullahy 2015).

Will had three other characteristics that we grew to appreciate, if not emulate, over the years. He was absolutely meticulous about research—data, methods, and attribution. Precision is not merely an abstract statistical concept but an essential part of all steps in a research project.

Will was extraordinarily generous with his time. We know of many junior economists who were amazed and profoundly grateful that Will took the time to give detailed feedback on their article or presentation—and to explain why their standard errors were all wrong.

Finally, Will was hilariously funny. We know this is a rare trait in an economist, but a little levity makes the daily give and take in search of truth that much more enjoyable.

We dedicate this book to our friend and colleague, Will Manning.

Partha Deb and Edward C. Norton

Contents

(Pages omitted)

Preface

This book grew out of our experience giving presentations about applied health econometrics at the International Health Economics Association and the American Society of Health Economists biennial conferences. In those preconference seminars, we tried to expose graduate students and early career academics to topics not generally covered in traditional econometrics courses but nonetheless are salient to most applied research on healthcare expenditures and use. Participants began to encourage us to turn our slides into a book.

In this book, we aim to provide a clear understanding of the most commonly used (and abused) econometric models for healthcare expenditure and use and of approaches to choose the most appropriate model. If you want intuition, meaningful examples, inspiration to improve your best practice, and enough math for rigor but not enough to cause rigor mortis, then keep reading. If you want a general econometrics textbook, then put down this book and go buy a general econometrics textbook. Get ready to try new methods and statistical tests in Stata as you read. Be prepared to think.

Despite years of training and practice in applied econometrics, we still learned a tremendous amount while working on this book from reading recent literature, comparing and testing models in Stata, and debating with each other. We particularly learned from our coauthor Will Manning, who unfortunately died in 2014 before seeing our collective effort come to fruition. Will was a fountain of knowledge. We think that his overarching approach to econometrics of repeated testing to find the best model for the particular research question and dataset is the best guide. The journey matters, not just the final parameter estimate.

In closing, we want to thank some of the many people who have helped us complete this book. David Drukker, editor and econometrician, had numerous suggestions, large and small, that dramatically improved the book. We are grateful to Stephanie White, Adam Crawley, and David Culwell at StataCorp for help with LaTeX, editorial assistance, and production of the book. We thank Betsy Querna Cliff, Morris Hamilton, Jun Li, and Eden Volkov for reading early drafts and providing critical feedback. We thank the many conference participants who were the early guinea pigs for our efforts at clarity and instruction and especially those who gave us the initial motivation to undertake this book. Our wives, Erika Bach and Carolyn Norton, provided support and encouragement, especially during periods of low marginal productivity. Erika Manning cared for Will during his illness and tolerated lengthy phone calls at odd hours, and Will's bad puns at all hours.

Partha Deb and Edward C. Norton

(Pages omitted)

1 Introduction

Health and healthcare are central to society and economic activity. This observation extends beyond the large fraction of gross national product devoted to formal healthcare to the fact that health and healthcare affect each other and numerous other decisions. Health affects people's ability to engage in work and leisure, their probability of marriage, probability of living to a ripe old age, and how much they spend on healthcare. Healthcare affects health mostly for the better, although side effects and medical errors can have drastic consequences. The desire for better health motivates decisions about smoking, drinking, diet, and exercise over a lifetime. Therefore, it is important to understand the underlying causes of health and how health affects people's lives, including examining the determinants of healthcare expenditures and use.

Economic theory and policy motivate research questions on healthcare expenditures and use. The healthcare sector is one of the best areas to test economic theories of insurance, the behavior of nonprofit firms, principal-agent relationships, and the influence of peers. Moreover, governments are intimately involved with healthcare. For example, in the United States, the federal and state governments administer Medicare and Medicaid health insurance, the Veterans Administration provides healthcare to veterans, and the government regulates tobacco, alcohol, and prescription drugs. Understanding what drives healthcare expenditures and use is essential for policy. National domestic political debates often center around policies that aim to enhance public health, improve quality of healthcare, and ensure affordable access to health insurance. Health economists have much to offer by studying these issues.

The last few decades have also seen a proliferation of sophisticated statistical methods. Researchers now have many alternatives to ordinary least squares (OLS) to analyze data with dependent variables that are binary, count, or skewed. Researchers can adjust estimates to control for complex survey design and heteroskedasticity. There are classes of models [for example, generalized linear models (GLM)] and statistical methods (for example, maximum likelihood estimation and generalized method of moments) beyond least squares that provide powerful and unified approaches to fitting many complex models. Advances in computing power mean that researchers can estimate technically complex statistical models faster than ever. Stata (and other statistical software) allows researchers to use these models quickly and easily.

Like the people behind the statistics, data come in all shapes, sizes, and ages. Researchers collect population health and census data, episode-level claims data, survey data on households and on providers, and, more recently, individual biometric data—including genetic information. Datasets are often merged to generate richer information over time. The variety of data is dizzying.

The importance of the research and policy questions requires that we use the econometric models with care and that we think deeply about the correct interpretation. Faster computers do not obviate the need for thought.

In this book, we lay out the main statistical approaches and econometric models used to analyze healthcare expenditure and use data. We explain how to estimate and interpret the models using easy-to-follow examples. We include numerous references to the main theoretical and applied literature. We also discuss the strengths and weaknesses of the models we present. Knowing the limitations of models is as important as knowing when to appropriately use them. Most importantly, we demonstrate rigorous model testing methods. By following our approach, researchers can rigorously address research questions in health economics using a way that is tailored to their data.

1.1 Outline

This book is divided into three groups of chapters. The early chapters provide the background necessary to understand the rest of the book. Many empirical research questions aim to estimate treatment effects. Consequently, chapter 2 introduces the potential outcomes framework, which is useful for estimating and interpreting treatment effects. It also relates treatment effects to marginal and incremental effects in both linear and nonlinear models. Chapter 3 introduces the Medical Expenditure Panel Survey dataset, which is used throughout this book for illustrative examples. Chapter 4 illustrates how to estimate the average treatment effect, the treatment effect on the treated, and marginal and incremental effects for linear regression models. Chapter 4 also shows that misspecifications in OLS models can lead to inconsistent average effects. It also includes graphical and statistical tests for model specification to help decide between competing statistical models.

The core chapters describe the most prominent set of models used for healthcare expenditures and use, including those that explicitly deal with skewness, heteroskedasticity, log transformations, zeros, and count data. Chapter 5 presents GLMs as an alternative to OLS for modeling positive continuous outcomes. Generalized linear models are especially useful for skewed dependent variables and for heteroskedastic error terms. Although we argue that GLM provides a powerful set of models for health expenditure, we also lay out the popular log transformation model in chapter 6. Transforming a dependent variable by taking its natural logarithm is a widely used way to model skewed outcomes. Chapter 6 describes several versions that differ in their assumptions about heteroskedasticity and the distribution of the error term (normal or nonnormal). We show that interpretation can be complex, even though estimation is simple. Chapter 7 adds observations with outcomes equal to zero. Most health expenditure data have a substantial mass at zero, which makes models that explicitly account for zeros appealing. Here we describe and compare two-part and selection models. We explain the underlying assumptions behind the often misunderstood two-part model, and show how two-part models are superficially similar, yet strikingly different from selection models in fundamental ways. Chapter 8 moves away from continuous dependent variables to

count models. These models are essential for outcomes that are nonnegative integer valued, including counts of office visits, number of cigarettes smoked, and prescription drug use.

The book then shifts to more advanced topics. Chapter 9 presents four flexible approaches to modeling treatment-effect heterogeneity. Quantile regression allows response heterogeneity by level of the dependent variable. We describe basic quantile regressions and how to use those models to obtain quantile treatment effects. Next, we describe finite mixture models. These models allow us to draw the sample from a finite number of subpopulations with different relationships between outcomes and predictors in each subpopulation. Thus, finite mixture models can uncover patterns in the data caused by heterogeneous types. Third, we describe local-linear regression, a nonparametric regression method. Nonparametric regression techniques make few assumptions about the functional form of the relationship between the outcome and the covariates and allow for very general relationships. Finally, conditional density estimation is another flexible alternative to linear models for dependent variables with unusual distributions. The last two chapters discuss issues that cut across all models. Chapter 10 introduces controlling for endogeneity or selection-on-unobservables of covariates of policy interest to the researcher. Chapter 11 discusses design effects. Many datasets have information collected with complex survey designs. Analyses of such data should account for stratified sampling, primary sampling units, and clustered data.

This book does not attempt to provide a comprehensive treatment of econometrics. For that, we refer readers to other sources (for example, Cameron and Trivedi [2005; 2010], Greene [2012], Wooldridge [2010; 2016]). Instead, we focus on healthcare econometric models that emphasize three core statistical issues of skewness, zeros, and heterogeneous response. We focus on providing intuition, a basic mathematical framework, and user-friendly Stata applications. We provide citations to the literature for original proofs and important applications. Much promising theoretical and applied work continues to appear in the literature each year. Jones (2010) describe some of the recent research as well as a wide range of econometric approaches.

1.2 Themes

Although we present numerous alternative models and ways to check and choose between those models, it should be no surprise that we do not determine a single best model for all situations or a good second-best model for all cases. Instead, researchers must find the model that is most appropriate for their research question and data. We recommend comprehensive model checking, but model checking is not a simple check list. It requires thought.

We aim to provide the tools to find the best model to consistently estimate the answer to the research question. This answer will often be a function of $E(y|\mathbf{x})$, such as the average treatment effect, or the marginal effect of a covariate on the outcome. We are also concerned about the precision of those estimates, measured by the variance of the estimators.

Of all possible statistical models, we focus on those that address three key issues that often appear in health expenditure and use data: skewness, a large mass at zero, and heterogeneous response. Health expenditure data are often wildly right skewed. Transforming the dependent variable to generate a dependent variable with a more symmetric distribution may improve the statistical properties of the model fit but may make it harder to interpret. The distributions of many interesting health outcomes—such as total annual healthcare expenditures, hospital visits in the calendar year, and smoking in the last 30 days—typically have a substantial fraction of zeros, which can pose difficulties for standard statistical models. Consequently, health economists have developed models to deal with such outcomes, allowing for a rich understanding of how variables affect whether the outcome is positive (extensive margin) and the magnitude of the outcome (intensive margin). While a single, summary marginal effect is sometimes of interest, we often expect heterogeneous treatment effects across different subpopulations. Modeling the heterogeneity explicitly can reveal new insights.

In summary, our aim is to improve best practices among health economists and health services researchers.

1.3 Health econometric myths

Despite the tremendous recent advances in econometrics, we have noticed a number of misconceptions in the published literature. We hope the following myths will disappear in future generations:

1. *Model selection by citation is safe.* The lemming approach to econometrics is to follow blindly what others have done. But each research question, each dataset, and each model requires individual attention. We advocate artisanal handcrafted research, not mass-produced cookie-cutter research (see chapter 2).
2. *Trim outliers.* Outliers are so annoying. They are highly influential, do not fit nicely onto graphs, and are just, well, different. Why not trim them from the data? The reason is that each outlier represents a real person or episode. As much as a hospital administrator would like to assume away an ultraexpensive patient, the patient exists and is an important feature of many datasets. Embrace the diversity; start by exploring outliers in the Medical Expenditure Panel Survey data described in chapter 3.
3. *OLS is fine.* OLS regression has many virtues. It is easy to estimate and interpret. Under a set of well-known assumptions—including that the model as specified is correct—OLS is the best linear unbiased estimator, except when the assumptions fail, which is often. We demonstrate the limitations of OLS in chapter 4.
4. *All GLM models should have a log link with a gamma distribution.* Several early influential articles using GLM models in health economics happened to analyze data for which the log link with a gamma distribution was the appropriate choice. Different link and distributional families may be better (see chapter 5) for other data.

5. *Log all skewed dependent variables.* Health economists have developed a compulsive, almost Pavlovian, instinct to log any and all skewed dependent variables. While the log transformation makes estimation on the log scale simple, it makes interpretation and prediction on the raw scale surprisingly difficult (see chapter 6).
6. *Use selection models for data with a large mass at zero.* When the data have substantial mass at zero, some researchers reach for the two-part model, while others reach for selection models. Their choices often lead to considerable argument over which is better. We advocate the two-part model for researchers interested in actual outcomes (including the zeros), and we advocate selection models for researchers interested in latent outcomes (assuming that the zeros are missing values). We set the record straight in chapter 7.
7. *All count models are Poisson.* Ever wonder why some researchers reflexively use Poisson, and others use the negative binomial? We explain the tradeoff between inference about the conditional mean function and conditional frequencies while providing intuition and pretty pictures (see chapter 8).
8. *Modeling heterogeneity is not worth the effort.* Tired of assuming monolithic treatment effects? Want to spice up your research life? We introduce four ways to model treatment-effect heterogeneity that can enrich any analysis (see chapter 9).
9. *Correlation is causation.* Actually, virtually all researchers know that statement is false. However, knowing it and correctly adjusting for endogeneity are two different things. We discuss ways to better establish causality by controlling for endogeneity, which is pervasive in applied social science research (see chapter 10).
10. *Complex survey design is just for summary statistics.* Most large surveys use stratification and cluster sampling to better represent important subsamples and to use resources efficiently. Model estimation, not just summary statistics, should control for sample weights, clustering, and stratification (see chapter 11).

1.4 Stata friendly

We assume the reader has a basic understanding of Stata. To learn more, read the Stata manuals and online help, or consult introductions to Stata by Long and Freese (2014) (see especially chapters 2–4) and Cameron and Trivedi (2010) (see chapters 1 and 2 for overview). Stata is easy to learn, easy to use, and has a powerful set of tools. Many of them are built-in, but others are provided by dedicated users who share their code via Stata packages. Once the reader has a grasp of the basics, our book will be fully accessible.

Merely reading about econometrics is not the best way to learn. Readers must actively analyze data themselves. Therefore, we provide user-friendly Stata code, so interested readers cannot only reproduce all the examples in the book but also modify the code to analyze their own data. The data and Stata code in this book are publicly available. We have designed this not only to be user-friendly but also to be interactive. Dig in!

1.5 A useful way forward

Finally, we agree with the observation by Box and Draper (1987) that "all models are wrong, but some are useful". Our intent is to provide methods to choose models that are useful for the research question of interest.

(Pages omitted)

4 The linear regression model: Specification and checks

4.1 Introduction

The linear regression model is undoubtedly the workhorse of empirical research. Researchers use it ubiquitously for continuous outcomes and often for count and binary outcomes. With relatively few assumptions—namely, the relationship between the outcome and the regressors is correctly specified, and the error term has an expected value of zero conditional on the values of the regressors—ordinary least-squares (OLS) estimates of the parameters of the model are unbiased and consistent. In other words, given those two assumptions, OLS delivers estimates that are correct on average. In addition, if the distribution of the errors have constant variance across the sample observations and are uncorrelated across sample observations, then OLS produces estimates that have the smallest variance among all linear unbiased estimators.

The formal statement of these properties is the Gauss–Markov theorem. Many textbooks formally discuss the assumptions and proof, including Wooldridge (2010) and Cameron and Trivedi (2005). The Gauss–Markov theorem has two main implications. First, OLS estimates have the desirable property of being unbiased under relatively weak conditions. Second, there is no linear estimator with better properties than OLS. These desirable features mean, in many cases, we can use the linear regression model to estimate causal treatment effects and marginal and incremental effects of other covariates, as we outlined in chapter 2. In this chapter, we show how these can be implemented in Stata and discuss the interpretation of various effects.

The Gauss–Markov theorem applies to OLS models only when the assumptions are met. If a regressor is endogenous, for example, the conditional expectation of the error term is not zero. If it was, it would violate one of the Gauss–Markov theorem's assumptions, and the OLS estimates would be inconsistent. With observational data, researchers should always be aware of the possibility of endogenous regressors. We address these issues in chapter 10.

The other main assumption is that the model specification is correct. Estimation of a linear model without serious consideration of the model specification can lead to substantially misleading answers. One of the most important features of any model is the relationship of the covariates to the dependent variable. Correct specification of the relationship is a key assumption of the theorem. In practice, while researchers cannot claim to know the true model, they should strive to specify good models. A

good model includes all the necessary variables—including higher-order polynomials and interactions terms—but no more. A good model includes variables with the correct functional relationship between the covariates and outcome. Choosing the correct model specification requires making choices. There is tension between simplicity and attention to detail, and there is tension between misspecification and overfitting. We address these issues in this chapter.

In this chapter, we show with two examples how easy it is to estimate inconsistent marginal effects when the fit model is misspecified. Marginal effects are surprisingly sensitive to model misspecification. If we include a variable in the model, but the relationship between it and the dependent variable is not correct, the estimated marginal effects of that variable are sensitive to the distribution of that covariate and to whether the marginal effects are conditioned on a specific value of that covariate.

Some readers may wonder why we obsess about model specification. A commonly held belief is that the estimate, $\widehat{\beta}_k$, of the average marginal effect (AME), $\partial E(y_i|\mathbf{x}_i)/\partial x_k$, of covariate, x_k, is consistent even if it is estimated using a misspecified model. We show that this can easily be false. In addition, we believe that the focus on average effects is too narrow a view, because policy interest is often about the response to a covariate for a specific value of that covariate. For example, we may care only about the effect of a weight-loss drug on those with an unusually high body mass index, rather than the entire population. In the case of health insurance, we might be worried about the effect of raising the coinsurance rates or deductibles in the least generous health insurance plans rather than for all health insurance plans. In such situations, the marginal effect for the subsample of interest may be inconsistent, even if the average of marginal effects for the full sample are not.

Because we never know the correct model specification (theory rarely provides guidance for model specification), it is important to know how to make informed choices. To this end, the final sections in this chapter describe visual and statistical methods to test model specification.

4.2 The linear regression model

It is useful to begin with a precise, mathematical formulation of the linear regression model, in which

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

where y_i is the outcome for the ith observation ($i = 1, \ldots, N$), $\mathbf{x}_i'$ is a row vector of k covariates including a constant, $\boldsymbol{\beta}$ is a column vector of k coefficients to be estimated including the intercept, and u is the error term. A linear specification can include nonlinear terms in $\mathbf{x}_i$ but is always linear in $\boldsymbol{\beta}$. Specifications that are nonlinear in $\boldsymbol{\beta}$ generally cannot be transformed into a linear specification.

As we showed in chapter 2, if the model is linear in variables, then the estimates of treatment, marginal, and incremental effects are all simply regression coefficients. Nonlinear terms, which are interactions between covariates or polynomial terms of covari-

ates, are functions of parameters and covariate values. We must estimate and interpret them more carefully.

4.3 Marginal, incremental, and treatment effects

We begin with a fairly simple OLS regression model to predict total annual expenditures at the individual level for those who spend at least some money, using the 2004 Medical Expenditure Panel Survey (MEPS) data (see chapter 3). Our goal is to interpret the results using the framework of potential outcomes and marginal effects (see chapter 2). To be clear, the model we fit may not be appropriate for a serious research exercise: it drops all observations with zero expenditures, and its specification of covariates is rudimentary. Additionally, we do not consider any possible models besides OLS, especially ones that may be better suited to deal with the severe skewness in the distribution of this outcome, and we do not control for design effects or possible endogeneity. In short, we ignore all interesting features of this typical health outcome variable, knowing that we will return to each of these issues throughout the book. The focus of this section is to provide a framework for interpreting regression results.

In this regression model, we estimate the effect of age (`age`), gender (`female` is a binary indicator for being female), and any health limitations (`anylim` is a binary indicator of whether the person has health limitations) on total healthcare expenditures for persons with any expenditures (`exptot` > 0), using the MEPS data (see chapter 3).

```
. *** MEPS data
. use http://www.stata-press.com/data/heus/heus_mepssample
(Sample of MEPS 2004 data)
. *** Restrict to subsample with positive total expenditures
. drop if exp_tot <= 0
(3,440 observations deleted)
```

We include an interaction term between age and gender, allowing the effect of age to differ between men and women. It is essential to use the notation with `##` between `c.age` and `i.female` so Stata understands that those variables are interacted. The prefix `c.` indicates that the variable age is continuous; the prefix `i.` indicates that the variable gender is binary. We estimate robust standard errors.

The results appear to show that healthcare expenditures increase with age and are higher for women. However, the interaction term is negative and statistically significant, indicating that we must put more effort into fully understanding the relationship between these demographics and total expenditures. Unsurprisingly, expenditures are far higher for those with at least one limitation. All coefficients are statistically significant at $p < 0.05$.

```
. *** Regression to predict total expenditures
. regress exp_tot c.age##i.female i.anylim, vce(robust)
Linear regression                               Number of obs     =     15,946
                                                F(4, 15941)       =     221.41
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0773
                                                Root MSE          =      10187
```

exp_tot	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	94.70939	8.664488	10.93	0.000	77.72601	111.6928
female						
Female	1593.099	459.1437	3.47	0.001	693.1256	2493.073
female#c.age						
Female	-22.56145	10.64236	-2.12	0.034	-43.42168	-1.701215
anylim						
Activity ..	4456.928	237.8453	18.74	0.000	3990.724	4923.132
_cons	-1772.051	380.9481	-4.65	0.000	-2518.752	-1025.349

4.3.1 Marginal and incremental effects

We first interpret the results for age and gender in more detail. Following chapter 2, we interpret regression results for the continuous variable `age` as a marginal effect (derivative) and for the dichotomous variable `female` as an incremental effect (difference). One way to interpret the effects (not necessarily the most informative way for this example, as we will see) is to compute the average marginal and incremental effects using the Stata command `margins, dydx()`. Because of the interaction term between age and gender, the average marginal and incremental effects will not equal any of the estimated coefficients in the model.

Women spend more than men by an average of $523, averaged across all ages in the sample. The AME of `age` is $81, meaning that on average (allowing for the interaction with `female`) for this sample, an increase in age of 1 year corresponds with an increase in total expenditures of $81. We note that this interpretation, in a model with covariates entered nonlinearly, is only true because the model is affine in `age`. If one is interested in knowing what happens if each person in the sample becomes one year older, it would be better to do that computation directly. We show how this can be done using `margins` in section 5.7.

```
. *** Average marginal effects for age and female
. margins, dydx(age female)
Average marginal effects                        Number of obs     =     15,946
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age 1.female
```

	Delta-method dy/dx	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	81.38701	5.338745	15.24	0.000	70.92247	91.85156
female						
Female	523.2414	171.3697	3.05	0.002	187.3375	859.1454

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

The average marginal and incremental effects calculated with `margins, dydx()` do not fully illustrate the complex relationship between age, gender, and total expenditures because of the interaction between them. One way to improve the interpretation is to calculate the marginal effect of age separately for men and for women (using the `at(female=(0 1))` option). The marginal effect of age for men is about $23 higher than for women ($95 compared with $72). This means that as men age, their spending increases faster than that of women.

```
. *** Marginal effect of age by gender
. margins, dydx(age) at(female=(0 1))
Average marginal effects                        Number of obs     =     15,946
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age
1._at        : female          =           0
2._at        : female          =           1
```

	Delta-method dy/dx	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age						
_at						
1	94.70939	8.664488	10.93	0.000	77.72601	111.6928
2	72.14794	6.52578	11.06	0.000	59.35668	84.93921

Similarly, we calculate the incremental effect of gender at different ages (using the `at(age=(20 45 70))` option). The incremental effect of gender is different at each age, being more than $1,140 at age 20, and close to 0 around age 70.

(Pages omitted)

5.2.2 Parameter estimation

As alluded to above, GLM estimation requires two sets of choices. The first set of choices determines the link function and the distribution family. In section 5.8, we discuss how we chose based on rigorous statistical tests.

For the parameter estimates in the model to be consistent, it is only necessary to correctly specify the link function, g, and how the covariates enter the index function. The choice of the distribution family affects the efficiency of the estimates, but an incorrect choice does not lead to inconsistency of the parameter estimates. An inappropriate assumption about the distribution family can lead to an inconsistent estimate of the inference statistics, but this inconsistency can be remedied using robust standard errors.

The other choice is whether to estimate GLMs by quasi–maximum likelihood or iteratively reweighted least squares. In Stata, the default is quasi–maximum likelihood, which does not require specification of the full log likelihood. The choice between these two methods does not seem to matter much in practice for typical models and datasets.

After fitting a GLM, one can easily derive marginal and incremental effects of specific covariates on the expected value of y (or other treatment effects).

5.3 GLM examples

We now show how to estimate GLMs for healthcare expenditures with a few choices of link functions and distribution families, using the MEPS data introduced in chapter 3. Specifically, we estimate the effect of age (`age`) and gender (`female` is a binary indicator for being female) on total healthcare expenditures for persons with any expenditures (`exp_tot` > 0).

Our first example is a model with a log link (option `link(log)`) and a Gaussian family (option `family(gaussian)`). This is equivalent to a nonlinear regression model with an exponential mean. The results show that healthcare expenditures increase with age and are higher for women, but the coefficient on `female` is not statistically significant at the 5% level. Because the conditional mean has an exponential form, coefficients can be interpreted directly as percent changes. Expenditures increase by about 2.6% with each additional year of age after adjusting for gender. Women spend about 8% more than men $[0.080 = \exp(0.0770) - 1]$ after controlling for age.

```
. *** GLM of total expenditures, log link and Gaussian family
. glm exp_tot age female, link(log) family(gaussian) vce(robust)
Iteration 0:   log pseudolikelihood = -191206.56
Iteration 1:   log pseudolikelihood = -183815.74
Iteration 2:   log pseudolikelihood = -172074.86
Iteration 3:   log pseudolikelihood = -170297.76
Iteration 4:   log pseudolikelihood = -170068.69
Iteration 5:   log pseudolikelihood = -170067.92
Iteration 6:   log pseudolikelihood = -170067.92
Generalized linear models                         No. of obs      =     15,946
Optimization     : ML                             Residual df     =     15,943
                                                  Scale parameter =   1.07e+08
Deviance         =  1.71351e+12                   (1/df) Deviance =   1.07e+08
Pearson          =  1.71351e+12                   (1/df) Pearson  =   1.07e+08
Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =   21.33086
Log pseudolikelihood = -170067.9162               BIC             =   1.71e+12
```

exp_tot	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0257328	.00101	25.48	0.000	.0237533	.0277123
female	.0769791	.0420732	1.83	0.067	-.0054828	.1594411
_cons	7.041415	.0652623	107.89	0.000	6.913503	7.169327

The second example also specifies a log link but assumes that the distribution family is gamma (option `family(gamma)`), implying that the variance of expenditures is proportional to the square of the mean. This is a leading choice in published models of healthcare expenditures, but we will return to the choices of link and family more comprehensively in section 5.8.

The results show that healthcare expenditures increase with age and are higher for women. Both coefficients are statistically significant, with $p < 0.001$. Expenditures increase by about 2.8% with each additional year of age, which is quite close to the effect fit by the model with the Gaussian family. However, now we find that women spend about 23% more than men ($0.23 = \exp(0.2086) - 1$), after controlling for age. This is almost three times as large as the effect estimated in the model with the Gaussian family. A small change in the model leads to a large change in interpretation.

```
. *** GLM of total expenditures, log link and gamma family
. glm exp_tot age female, link(log) family(gamma) vce(robust)
Iteration 0:   log pseudolikelihood = -150164.47
Iteration 1:   log pseudolikelihood = -148063.58
Iteration 2:   log pseudolikelihood =  -148047.8
Iteration 3:   log pseudolikelihood = -148047.79
Generalized linear models                         No. of obs      =     15,946
Optimization     : ML                             Residual df     =     15,943
                                                  Scale parameter =   7.097257
Deviance         =  33478.20882                   (1/df) Deviance =   2.099869
Pearson          =  113151.5612                   (1/df) Pearson  =   7.097257
Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =   18.56902
Log pseudolikelihood =  -148047.791               BIC             =  -120801.6
```

exp_tot	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	.0279516	.0012071	23.16	0.000	.0255856	.0303176
female	.2086064	.0478756	4.36	0.000	.114772	.3024408
_cons	6.835683	.0856778	79.78	0.000	6.667757	7.003608

Our primary intent was to use these examples to demonstrate the use of the `glm` command and explain how to interpret coefficients. However, these examples also show that the estimated effects in a sample can be quite different across distribution family choices when the link function is the same, even though the choice of family has no theoretical implications for consistency of parameter estimates.

We could run many other GLM models, changing the link function or the distributional family. For example, we could fit a GLM with a square root link (option `link(power 0.5)`) and a Poisson family (option `family(poisson)`). Or we could fit a GLM with a cube root link (option `link(power 0.333)`) and an inverse Gaussian family (option `family(igaussian)`).

5.4 GLM predictions

For all GLM models with a log link, the expected value of the dependent variable, y, is the exponentiated linear index function:

$$E\left(y_i|\mathbf{x}_i\right) = \mu_i = g^{-1}\left(\mathbf{x}_i'\boldsymbol{\beta}\right) = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right) \tag{5.1}$$

The sample average of the expected value of total expenditures is the average of μ_i over the sample. We calculate its estimate using the `margins` command. The predicted mean of total expenditures is \$4,509, less than 1% from the sample mean of \$4,480.

```
. *** Predicted mean of total expenditures from GLM with log link and gamma
> family
. quietly glm exp_tot age female, link(log) family(gamma) vce(robust)
. margins
Predictive margins                              Number of obs     =     15,946
Model VCE    : Robust
Expression   : Predicted mean exp_tot, predict()
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	4508.963	81.41072	55.39	0.000	4349.401	4668.525

When we compare predictions from log transformation models in chapter 6 with the sample mean, we will find that those predictions are much further off. They will be anywhere from 10% to 20% too high. GLM is generally better than log models at reproducing the sample mean of the outcome.

5.5 GLM example with interaction term

Before computing marginal effects, we extend our simple specification to include the interaction of age and gender as a covariate. That is, we allow for the effect of gender to vary by age (or equivalently, the effect of age to vary by gender). The results with an interaction term are harder to interpret, but more realistic, and will help show the power of several Stata postestimation commands.

When including interaction terms, one must use special Stata notation, so that `margins` knows the relationship between variables when it takes derivatives. Therefore, we use `c.` as a prefix to indicate that `age` is a continuous variable, `i.` to indicate that `female` is an indicator variable, and `##` between them to include not only the main effects but also their interaction.