Microeconometrics Using Stata

Volume I: Cross-Sectional and Panel Regression Methods

Second Edition

A. COLIN CAMERON
Department of Economics
University of California, Davis, CA
and
School of Economics
University of Sydney, Sydney, Australia

PRAVIN K. TRIVEDI
School of Economics
University of Queensland, Brisbane, Australia
and
Department of Economics
Indiana University, Bloomington, IN

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: 1-59718-359-8 (volumes I and II)
Print ISBN-10: 1-59718-361-X (volume I)
Print ISBN-10: 1-59718-362-8 (volume II)
Print ISBN-13: 978-1-59718-359-8 (volumes I and II)
Print ISBN-13: 978-1-59718-361-1 (volume I)
Print ISBN-13: 978-1-59718-362-8 (volume II)
ePub ISBN-10: 1-59718-360-1 (volumes I and II)
ePub ISBN-10: 1-59718-363-6 (volumes I)
ePub ISBN-10: 1-59718-364-4 (volumes II)
ePub ISBN-13: 978-1-59718-360-4 (volumes I and II)
ePub ISBN-13: 978-1-59718-363-5 (volumes I)
ePub ISBN-13: 978-1-59718-364-2 (volumes II)

Library of Congress Control Number: 2022938057

Microeconometrics Using Stata

Volume II: Nonlinear Models and Causal Inference Methods

Second Edition

A. COLIN CAMERON
Department of Economics
University of California, Davis, CA
and
School of Economics
University of Sydney, Sydney, Australia

PRAVIN K. TRIVEDI
School of Economics
University of Queensland, Brisbane, Australia
and
Department of Economics
Indiana University, Bloomington, IN

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: 1-59718-359-8 (volumes I and II)
Print ISBN-10: 1-59718-361-X (volume I)
Print ISBN-10: 1-59718-362-8 (volume II)
Print ISBN-13: 978-1-59718-359-8 (volumes I and II)
Print ISBN-13: 978-1-59718-361-1 (volume I)
Print ISBN-13: 978-1-59718-362-8 (volume II)
ePub ISBN-10: 1-59718-360-1 (volumes I and II)
ePub ISBN-10: 1-59718-363-6 (volumes I)
ePub ISBN-10: 1-59718-364-4 (volumes II)
ePub ISBN-13: 978-1-59718-360-4 (volumes I and II)
ePub ISBN-13: 978-1-59718-363-5 (volumes I)
ePub ISBN-13: 978-1-59718-364-2 (volumes II)

Library of Congress Control Number: 2022938057

Contents

Contents

(Pages omitted)

Preface to the Second Edition

Microeconometrics Using Stata, published in December 2008, was written for Stata 10.1. *Microeconometrics Using Stata, Revised Edition*, published in January 2010, was written for Stata 11.0. This second edition is written for Stata 17.

Whereas the scope and coverage of the preceding editions were reasonably synchronized with our own *Microeconometrics: Methods and Applications* (Cambridge, 2005), this second edition has broader scope in several respects. We have attempted not only to update our previous coverage to bring it in line with newer tools in the latest edition of Stata but also to bring into the book many topics and methods that are now actively studied and increasingly used in applied microeconometrics. This coverage includes several topics, listed below, that were not covered in our 2005 text.

This second edition covers over ten years of both enhancements to Stata and developments in the methods most commonly used in empirical microeconometrics analysis. The focus of the book remains the use of linear and nonlinear regression methods for cross-sectional and short panel data. In particular, we give only short treatment to other features of Stata that are useful for data analysis such as data management, use within Stata of other programming languages such as Python, and automated document preparation. The new edition is much expanded and is split into two volumes.

The first volume, comprising chapters 1–15 and Stata and Mata appendixes, focuses on the linear regression model and provides a brief introduction to nonlinear regression models. This volume is an expanded version of chapters 1–10, 12–13, and the appendixes of the first and revised editions. In places, there is greater explanation of underlying methods, and much of the first volume is intended to be suitable for an advanced undergraduate course in addition to serving graduate students and researchers.

The second volume, comprising chapters 16–30, covers the standard nonlinear models as well as more advanced and more recent material. In addition to updated versions of chapters 14–18 of the first edition and the revised edition, the second volume includes new chapters on duration models, treatment effects in randomized control trials, treatment effects with endogenous treatments, parametric models for endogeneity and heterogeneity, spatial regression, semiparametric regression, machine learning and prediction, and Bayesian methods.

Some methods we cover are well established. Other methods we present are in areas of active research, so they may become replaced by better methods. In particular, many methods for causal analysis using observational or experimental methods are still being established and improved upon, at a remarkably rapid pace. This includes

inference for instrumental variables with weak instruments, cluster–robust inference with few clusters, treatment-effects estimation with heterogeneous treatment effects, regression discontinuity design, and causal analysis using machine learning methods. Accordingly, we plan to periodically add some supplementary material on the book's website (http://cameron.econ.ucdavis.edu/mus2).

Our target user base consists of practitioners of applied microeconometrics. This group is quite diverse in terms of familiarity with the available econometric tools. In deference to such diversity, we have chosen to separate the more advanced aspects of many topics and place them in different parts of the book. This is a challenging task because often the same material could, and in some cases should, appear in several alternative places. To assist the reader, we have provided numerous cross-references and a much lengthier subject index. The reader will benefit from checking out these connections.

Datasets and the do-files used in this book are available on the Stata Press website at https://www.stata-press.com/data/mus2.html. Any corrections to the book will be documented at https://www.stata-press.com/books/microeconometrics-stata/.

The preparation of this second edition has benefited from generous help from many sources. We thank our colleagues, coauthors, students, and many users of the previous editions for their suggested improvements, for reading parts of the book, for permission to use datasets developed in joint research, and for encouragement to proceed with the project. We have benefited from presenting some of the material in various short courses around the world and from positive feedback from readers of the earlier editions that encouraged writing this updated edition. Colin Cameron would especially like to thank Shu Shen, Takuya Ura, Oscar Jorda, Marianne Bitler, the broader econometrics and empirical microeconomics community at the University of California–Davis, and Doug Miller and Adrian Pagan. Pravin Trivedi gratefully acknowledges the support provided by the School of Economics, University of Queensland. We thank Yulia Marchenko and Nikolay Balov for very detailed comments on the Bayesian chapters, and Kristin MacDonald for a careful reading of the final draft of the book. We thank David Culwell for his excellent editing and Stephanie White for managing the LaTeX formatting and production of this book. Most especially, both authors acknowledge their debt of gratitude to David Drukker for extensive feedback on many aspects of the material in the book throughout this project, including a complete reading, as well as feedback on the substantive aspects of applying the econometric and statistical tools. Finally, we thank our respective families for their patience and understanding during the long gestation period of the evolution of this project.

Davis, CA
Charlottesville, VA
June 2022

A. Colin Cameron
Pravin K. Trivedi

(Pages omitted)

3 Linear regression basics

3.1 Introduction

Linear regression analysis is often the starting point of an empirical investigation. Because of its relative simplicity, it is useful for illustrating the different steps of a typical modeling cycle that involves an initial specification of the model followed by estimation, diagnostic checks, and model respecification. The purpose of such a linear regression analysis may be to summarize the data, generate conditional predictions, or test and evaluate the role of specific regressors. We will illustrate these aspects using a specific data example.

This chapter is limited to basic linear regression analysis on cross-sectional data of a continuous dependent variable. The setup is for a single equation and exogenous regressors. Some standard complications of linear regression, such as misspecification of the conditional mean and model errors that are heteroskedastic, will be considered. In particular, we model the natural logarithm of medical expenditures instead of the level. We will ignore other various aspects of the data that can lead to more sophisticated nonlinear models presented in later chapters.

3.2 Data and data summary

The first step is to decide what dataset will be used. In turn, this decision depends on the population of interest and the research question itself. We discussed how to convert a raw dataset to a form amenable to regression analysis in section 2.4. In this section, we present ways to summarize and gain some understanding of the data, a necessary step before any regression analysis.

3.2.1 Data description

We analyze medical expenditures in 2003 of individuals 65 years and older who qualify for healthcare under the U.S. Medicare program. The original data source is the Medical Expenditure Panel Survey.

Medicare does not cover all medical expenses. For example, copayments for medical services and expenses of prescribed pharmaceutical drugs were not covered for the time period studied here. About half of eligible individuals therefore purchase supplementary insurance in the private market that provides insurance coverage against various out-of-pocket expenses.

In this chapter, we consider the impact of this supplementary insurance on total annual medical expenditures of an individual, measured in dollars. A formal investigation must control for the influence of other factors that also determine individual medical expenditure, notably, sociodemographic factors such as age, gender, education and income, geographical location, and health-status measures such as self-assessed health and presence of chronic or limiting conditions. In this chapter, as in other chapters, we instead deliberately use a short list of regressors. This permits shorter output and simpler discussion of the results, an advantage because our intention is to simply explain the methods and tools available in Stata.

3.2.2 Variable description

Given the Stata dataset for analysis, we begin by using the `describe` command to list various features of the variables to be used in the linear regression. The command without a variable list describes all the variables in the dataset. Here we restrict attention to the variables used in this chapter.

```
. * Variable description for medical expenditure dataset
. use mus203mepsmedexp
(A.C.Cameron & P.K.Trivedi (2022): Microeconometrics Using Stata, 2e)
. describe totexp ltotexp posexp suppins phylim actlim totchr age female income
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
totexp          double  %12.0g                Total medical expenditure
ltotexp         float   %9.0g                 ln(totexp) if totexp > 0
posexp          float   %9.0g      posexp     Total expenditure > 0
suppins         float   %9.0g      suppins    Has supp priv insurance
phylim          double  %12.0g     phylim     Has functional limitation
actlim          double  %12.0g     actlim     Has activity limitation
totchr          double  %12.0g                # of chronic problems
age             double  %12.0g                Age
female          double  %12.0g     female     Female
income          double  %12.0g                Annual household income/1000
```

The variable types and format columns indicate that all the data are numeric. In this case, some variables are stored in single precision (`float`) and some in double precision (`double`). From the variable labels, we expect `totexp` to be nonnegative; `ltotexp` to be missing if `totexp` equals 0; `posexp`, `suppins`, `phylim`, `actlim`, and `female` to be 0 or 1; `totchr` to be a nonnegative integer; `age` to be positive; and `income` to be nonnegative or positive. Note that the integer variables could have been stored much more compactly as `integer` or `byte`. The variable labels provide a short description that is helpful but may not fully describe the variable. For example, the key regressor `suppins` was created by aggregating across several types of private supplementary insurance.

3.2.3 Summary statistics

It is essential in any data analysis to first check the data by using the `summarize` command.

```
. * Summary statistics for medical expenditure dataset
. summarize totexp ltotexp posexp suppins phylim actlim totchr age female income
```

Variable	Obs	Mean	Std. dev.	Min	Max
totexp	3,064	7030.889	11852.75	0	125610
ltotexp	2,955	8.059866	1.367592	1.098612	11.74094
posexp	3,064	.9644256	.1852568	0	1
suppins	3,064	.5812663	.4934321	0	1
phylim	3,064	.4255875	.4945125	0	1
actlim	3,064	.2836162	.4508263	0	1
totchr	3,064	1.754243	1.307197	0	7
age	3,064	74.17167	6.372938	65	90
female	3,064	.5796345	.4936982	0	1
income	3,064	22.47472	22.53491	-1	312.46

On average, 96% of individuals incur medical expenditures during a year; 58% have supplementary insurance; 43% have functional limitations; 28% have activity limitations; and 58% are female because the elderly population is disproportionately female because of the greater longevity of women. The only variable to have missing data is `ltotexp`, the natural logarithm of `totexp`, which is missing for the $(3064 - 2955) = 109$ observations with `totexp` $= 0$.

All variables have the expected range, except that income is negative. To see how many observations on `income` are negative, we use the `tabulate` command, restricting attention to nonpositive observations to limit output.

```
. * Tabulate variable
. tabulate income if income <= 0
```

Annual household income/1000	Freq.	Percent	Cum.
-1	1	1.14	1.14
0	87	98.86	100.00
Total	88	100.00	

Only one observation is negative, and negative income is possible for income from self-employment or investment. We include the observation in the analysis here, though checking the original data source may be warranted.

Much of the subsequent regression analysis will drop the 109 observations with 0 medical expenditures, so in a research article, it would be best to report summary statistics without these observations.

3.2.4 More detailed summary statistics

Additional descriptive analysis of key variables, especially the dependent variable, is useful. For `totexp`, the level of medical expenditures, `summarize, detail` yields

```
. * Detailed summary statistics of a single variable
. summarize totexp, detail
                    Total medical expenditure
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0             0
 5%          112             0
10%          393             0       Obs                3,064
25%         1271             0       Sum of wgt.        3,064
50%       3134.5                     Mean           7030.889
                        Largest      Std. dev.      11852.75
75%         7151        104823
90%        17050        108256       Variance       1.40e+08
95%        27367        123611       Skewness       4.165058
99%        62346        125610       Kurtosis       26.26796
```

Medical expenditures vary greatly across individuals, with a standard deviation of 11,853, which is almost twice the mean. The median of 3,135 is much smaller than the mean of 7,031, reflecting the skewness of the data. For variable x, the skewness statistic is a scale-free measure of skewness that estimates $E[\{(x-\mu)/\sigma\}^3] = E\{(x-\mu)^3\}/\sigma^3$, the third central moment standardized by the cube of the standard deviation. The skewness is zero for symmetrically distributed data. The value here of 4.17 indicates considerable right skewness. The kurtosis statistic is an estimate of $E[\{(x-\mu)/\sigma\}^4] = E\{(x-\mu)^4\}/\sigma^4$, the fourth central moment standardized by the fourth power of the standard deviation. The reference value is 3, the value for normally distributed data. The much higher value here of 26.27 indicates that the tails are much thicker than those of a normal distribution. You can obtain additional summary statistics by using the `centile` command to obtain other percentiles and by using the `table` command, which is explained in section 3.2.6.

We conclude that the distribution of the dependent variable is considerably skewed and has thick tails. These complications often arise for commonly studied individual-level economic variables such as expenditures, income, earnings, wages, and house prices. It is possible that including regressors will eliminate the skewness, but in practice, much of the variation in the data will be left unexplained ($R^2 < 0.3$ is common for individual-level data), and skewness and excess kurtosis will remain.

Such skewed, thick-tailed data suggest a model with multiplicative errors instead of additive errors. A standard solution is to transform the dependent variable by taking the natural logarithm. Here this is complicated by the presence of 109 0-valued observations. We take the expedient approach of dropping the zero observations from analysis in either logs or levels. This should make little difference here because only 3.6% of the sample is then dropped. A better approach, using two-part or selection models, is covered in sections 19.5–19.7.

The output for `tabstat` in section 3.2.6 reveals that taking the natural logarithm for these data essentially eliminates the skewness and excess kurtosis.

The community-contributed `fsum` command (Wolfe 2002) is an enhancement of `summarize` that enables formatting the output and including additional information such as percentiles and variable labels. The community-contributed `outsum` command (Papps 2006) produces a text file of means and standard deviations for one or more subsets of the data, for example, one column for the full sample, one for a male subsample, and one for a female subsample.

3.2.5 Tables of frequencies

One-way tables can be created by using the `tabulate` command, presented in section 3.2.3, the `table` command, and the `tabstat` command. Two-way tables can also be created by using these commands.

For two-way tables of frequencies, only `table` produces clean output. For example,

```
. * Two-way table of frequencies
. table female totchr
```

	# of chronic problems								
	0	1	2	3	4	5	6	7	Total
Female									
No	239	415	323	201	82	23	4	1	1,288
Yes	313	466	493	305	140	46	11	2	1,776
Total	552	881	816	506	222	69	15	3	3,064

provides frequencies for a two-way tabulation of gender against the number of chronic conditions. The option `stat(percent)` provides percentages rather than frequencies.

The `tabulate` command can provide both row and column percentages. For example,

```
. * Two-way table with row and column percentages and Pearson chi-squared
. tabulate female suppins, row col chi2
```

Key
frequency *row percentage* *column percentage*

Female	Has supp priv insurance: No	Yes	Total
No	488	800	1,288
	37.89	62.11	100.00
	38.04	44.92	42.04
Yes	795	981	1,776
	44.76	55.24	100.00
	61.96	55.08	57.96
Total	1,283	1,781	3,064
	41.87	58.13	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 14.4991 Pr = 0.000

Comparing the row percentages for this sample, we see that while a woman is more likely to have supplemental insurance than not, the probability that a woman in this sample has purchased supplemental insurance is lower than the probability that a man in this sample has purchased supplemental insurance. Although we do not have the information to draw these inferences for the population, the results for Pearson's chi-squared test soundly reject the null hypothesis that these variables are independent. Other tests of association are available. The related command `tab2` will produce all possible two-way tables that can be obtained from a list of several variables.

For multiway tables, it is best to use `table`. For the example at hand, we have

```
. * Three-way table of frequencies
. table female suppins totchr, nototals
```

	# of chronic problems							
	0	1	2	3	4	5	6	7
Female								
No								
Has supp priv insurance								
No	102	165	121	68	25	6	1	
Yes	137	250	202	133	57	17	3	1
Yes								
Has supp priv insurance								
No	135	212	233	134	56	22	1	2
Yes	178	254	260	171	84	24	10	

An alternative is to use `tabulate` with the `by` prefix, but the results are not as neat as those from `table`.

3.2.6 Tables of summary statistics

The preceding tabulations will produce voluminous output if one of the variables being tabulated takes on many values. Then it is much better to use command `table` with the `statistics()` option to present tables that give key summary statistics for that variable, such as the mean and standard deviation. Note that the `statistics()` option, abbreviated `stat()`, was introduced in Stata 17 and replaces the `contents()` option available in earlier versions of Stata. Such tabulations can be useful even when variables take on few values. For example, when summarizing the number of chronic problems by gender, `table` yields

```
. * One-way table of summary statistics
. table (result) female, stat(count totchr) stat(mean totchr) stat(sd totchr)
>     stat(p50 totchr)
```

	Female		
	No	Yes	Total
Number of nonmissing values	1,288	1,776	3,064
Mean	1.659938	1.822635	1.754243
Standard deviation	1.261175	1.335776	1.307197
50th percentile	1	2	2

Women on average have more chronic problems (1.82 versus 1.66 for men). The option `stat()` can produce many other statistics, including the minimum, maximum, and key percentiles.

The `table` command with the `stat()` options can additionally produce two-way and multiway tables of summary statistics. As an example,

```
. * Two-way table of summary statistics
. table female suppins, stat(count totchr) stat(mean totchr) nototals
```

	Has supp priv insurance	
	No	Yes
Female		
No		
Number of nonmissing values	488	800
Mean	1.530738	1.73875
Yes		
Number of nonmissing values	795	981
Mean	1.803774	1.83792

shows that those with supplementary insurance on average have more chronic problems. This is especially so for males (1.74 versus 1.53).

The `tabulate, summarize()` command can be used to produce one-way and two-way tables with means, standard deviations, and frequencies. This is a small subset of the statistics that can be produced using `table`, so we might as well use `table`.

The `tabstat` command provides a table of summary statistics that permits more flexibility than `summarize`. The following output presents summary statistics on medical expenditures and the natural logarithm of expenditures that are useful in determining skewness and kurtosis.

```
. * Summary statistics obtained using command tabstat
. tabstat totexp ltotexp, statistics(count mean p50 sd skew kurt)
> columns(statistics)
```

Variable	N	Mean	p50	SD	Skewness	Kurtosis
totexp	3064	7030.889	3134.5	11852.75	4.165058	26.26796
ltotexp	2955	8.059866	8.111928	1.367592	-.3857887	3.842263

This reproduces information given in section 3.2.4 and shows that taking the natural logarithm eliminates most skewness and kurtosis. The `columns(statistics)` option presents the results with summary statistics being given in the columns and each variable being given in a separate row. Without this option, we would have summary statistics in rows and variables in the columns. A two-way table of summary statistics can be obtained by using the `by()` option.

The `collect` command, introduced in Stata 17, provides great flexibility in creating production-quality tables. The command is illustrated in section 3.5.7.

3.2.7 Hypothesis tests on the population mean

The `ttest` command can be used to test hypotheses about the population mean of a single variable (H_0: $\mu = \mu^*$ for specified value μ^*) and to test the equality of means (H_0: $\mu_1 = \mu_2$). For more general analysis of variance and analysis of covariance, the `oneway` and `anova` commands can be used, and several other tests exist for more specialized examples such as testing the equality of proportions.

These commands are rarely used in microeconometrics because they can be recast as a special case of regression with an intercept and appropriate indicator variables. Furthermore, regression has the advantage of reliance on less restrictive distributional assumptions, provided samples are large enough for asymptotic theory to provide a good approximation.

For examples of the `ttest` command and comparison with tests based on OLS estimation, see section 3.5.12.

3.2.8 Data plots

It is useful to plot a smoothed histogram or a density estimate of the dependent variable. Here we use the `kdensity` command, which provides a kernel estimate of the density.

The data are highly skewed, with a 97th percentile of approximately $40,000 and a maximum of $125,000. The `kdensity totexp` command will therefore bunch 97% of the density in the first 30% of the x axis. One possibility is to type `kdensity totexp if totexp < 40000`, but this produces a kernel density estimate assuming the data are truncated at $40,000. Instead, we use command `kdensity totexp`, we save the evaluation points in `kx1` and the kernel density estimates in `kd1`, and then we line-plot `kd1` against `kx1`.

We do this for both the level and the natural logarithm of medical expenditures, and we use `graph combine` to produce a figure that includes both density graphs (shown in figure 3.1). We have

```
. * Kernel density plots with adjustment for highly skewed data
. kdensity totexp if posexp==1, generate(kx1 kd1) n(500)
. graph twoway (line kd1 kx1) if kx1 < 40000, name(levels, replace)
. label variable ltotexp "Natural logarithm of expenditure"
. kdensity ltotexp if posexp==1, generate(kx2 kd2) n(500)
. graph twoway (line kd2 kx2) if kx2 < ln(40000), name(logs, replace)
. graph combine levels logs, iscale(1.2) ysize(2.5) xsize(6.0)
```

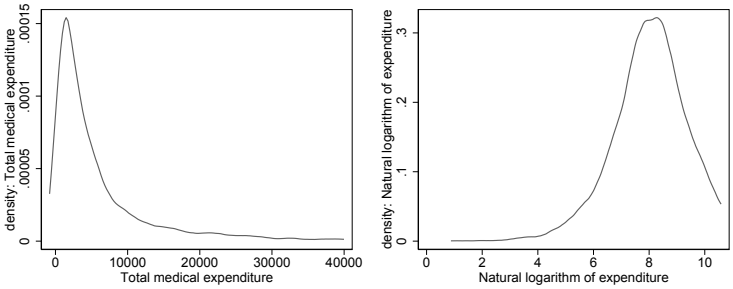

Figure 3.1. Comparison of densities of level and natural logarithm of medical expenditures

Only positive expenditures are considered, and for graph readability, the very long right tail of `totexp` has been truncated at $40,000. In figure 3.1, the distribution of `totexp` is very right skewed, whereas that of `ltotexp` is fairly symmetric.

3.3 Transformation of data before regression

When one specifies a linear regression model, the presumption is that the specified relationship between the variable of interest y and the regressors $\mathbf{x}$ is linear, which means that the marginal response of y to a unit change in x is constant.

The preferred model linking y and the regressors, however, may not be linear. For example, the relation between total production costs and output is usually specified to be nonlinear. In such cases, it is usual to interpret the regression as linear after transformation from the original units. Transformations to the linear form may involve both y and $\mathbf{x}$, or just one of those components.

The purpose of the transformation is to "straighten out" a relationship. Consider some leading examples. Suppose that the relationship takes the form $y = \exp(\beta_1 + \beta_2 x + u)$, where x denotes the regressor and u is the error term. Then the transformation $\ln y = \beta_1 + \beta_2 x + u$ is a "semilog" or "log-linear" regression that relates $\ln y$ to x. After transformation, β_2 measures $\partial E(\ln y)/\partial x = (1/y) \times \partial y/\partial x$, which varies inversely with y.

Now consider the multiplicative relationship $y = e^{\beta_1} x^{\beta_2} u$. Taking logs on both sides of the equality yields $\ln y = \beta_1 + \beta_2 \ln x + \ln u$, a linear-in-logs or log–log regression. In this case, the coefficient β_1 measures $\partial E\{\ln(y)/\partial \ln x\}$, that is, the elasticity of y with respect to x based on the constant elasticity model.

In both the preceding examples, while the dependent variable and, in the second example, the regressor have been transformed, the transformed models are linear in the parameters. So the transformed models can be fit using OLS regression, the subject of this chapter.

More generally, we may consider a regression such as $g(y) = f(x, \beta) + u$, where $g(\cdot)$ and $f(\cdot)$ denote some linearizing transformation whose specific form will depend upon the context. Choosing the functions that provide the best approximations to the data-generating process (DGP) is a part of model specification. Having chosen one, one relies on statistical tests to check whether the functional form is such that the remaining unexplained variation is roughly random.

Although the least-squares estimator of the linear regression requires only the data, or the error on the regression, to have quite weak distributional properties, transformations are often motivated by a preference for some particular features. For example, some outcomes such as income and expenditure often display a highly skewed distribution. A log transformation will typically make the distribution more symmetric and less nonnormal.

Another motivation for transformation is to make the error variance less heteroskedastic. For example, in its original form, a regression may display dependence between (say) the location parameter $E(y|x)$ and scale parameter $\text{Var}(y|x)$; a transformation may get rid of such dependence by reducing the heteroskedasticity of the error term. A family of power transformations, known as Box–Cox transformations, that replaces y by y^p is motivated by a similar consideration. A special case is $p = 1/2$, the square-root transformation. In a third example, suppose y is positive and we want to ensure that fitted values of y remain positive. A log transformation ensures this. In the final example, suppose y is a proportion, that is, $0 < y < 1$, and again we want the fitted values from the regression to preserve this property, whereas the linear regression of y on x will not. The logit transformation uses the transformed dependent variable $\log\{y/(1-y)\}$, which satisfies this requirement. This transformation also changes the range of values of the dependent variable, producing greater symmetry and spread in the tails of the distribution. In some cases, such changes make the least-squares estimator more robust.

Transformations generally affect the interpretations of regression coefficients, and transformations involving the dependent variables will also affect measures of goodness of fit such as regression R^2. This means that regression statistics such as R^2 with $g(y)$ as a dependent variable cannot be directly compared with those with y as the dependent variable. This complicates the comparison of regressions with different transformations of the dependent variable. A substantial literature exists on the topic of comparison of linear and linear-in-logs regressions; see Godfrey and Wickens (1981) and references cited there.

Finally, even if one chooses to regress $g(y)$ on $h(x)$, one may want to interpret the results in terms of the original units of y and x. This involves a thorny problem of *retransformation* that is discussed in section 4.2.3. In some cases, retransformation can

be avoided by directly modeling y using methods more advanced than OLS regression. In particular, we can use Poisson regression (`poisson` command) in place of the log-linear model and use logit regression (`logit` command) for proportions data.

Economic theory rarely suggests a specific parametric form of a regression model, thereby leaving room for empirical explorations. Nonparametric regressions (see section 14.6) are less restrictive in this respect.

3.4 Linear regression

We present the linear regression model, first in levels and then for a transformed dependent variable, here in logs.

3.4.1 Basic regression theory

We begin by introducing terminology used throughout the rest of this book. Let $\boldsymbol{\theta}$ denote the vector of parameters to be estimated, and let $\widehat{\boldsymbol{\theta}}$ denote an estimator of $\boldsymbol{\theta}$. Ideally, the distribution of $\widehat{\boldsymbol{\theta}}$ is centered on $\boldsymbol{\theta}$ with small variance, for precision, and a known distribution, to permit statistical inference. We restrict analysis to estimators that are consistent for $\boldsymbol{\theta}$, meaning that in infinitely large samples, $\widehat{\boldsymbol{\theta}}$ equals $\boldsymbol{\theta}$ aside from negligible random variation. This is denoted by $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ or, more formally, by $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0$ denotes the unknown "true" parameter value. A necessary condition for consistency is correct model specification or, in some leading cases, correct specification of key components of the model, most notably the conditional mean.

Under additional assumptions, most of the estimators considered in this book are asymptotically normally distributed, meaning that their distribution is well approximated by the multivariate normal in large samples. This is denoted by

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} N\left\{\boldsymbol{\theta},\ \text{Var}\left(\widehat{\boldsymbol{\theta}}\right)\right\}$$

where $\text{Var}(\widehat{\boldsymbol{\theta}})$ denotes the (asymptotic) variance–covariance matrix of the estimator (VCE). More efficient estimators have smaller VCEs. The VCE depends on unknown parameters, so we use an estimate of the VCE, denoted by $\widehat{V}(\widehat{\boldsymbol{\theta}})$. Standard errors of the parameter estimates are obtained as the square root of diagonal entries in $\widehat{V}(\widehat{\boldsymbol{\theta}})$. Different assumptions about the DGP, such as heteroskedasticity, can lead to different estimates of the VCE.

Test statistics based on asymptotic normal results lead to the use of the standard normal distribution and chi-squared distribution to compute critical values and p-values. For some estimators, notably, the OLS estimator, tests are instead based on the t distribution and the F distribution. This makes essentially no difference in large samples with, say, degrees of freedom greater than 100, but in practice it provides a better approximation especially for cluster–robust inference with few clusters; see section 3.4.6.

3.4.2 OLS regression and matrix algebra

The goal of linear regression is to estimate the parameters of the linear conditional mean

$$E(y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K \tag{3.1}$$

where usually an intercept is included so that $x_1 = 1$. Here $\mathbf{x}$ is a $K \times 1$ column vector with the jth entry—the jth regressor x_j—and $\boldsymbol{\beta}$ is a $K \times 1$ column vector with the jth entry β_j.

Sometimes, $E(y|\mathbf{x})$ is of direct interest for prediction. More often, however, econometrics studies are interested in one or more of the associated marginal effects (MEs),

$$\frac{\partial E(y|\mathbf{x})}{\partial x_j} = \beta_j$$

for the jth regressor. For example, we are interested in the MEs of supplementary private health insurance on medical expenditures. An attraction of the linear model is that estimated MEs are given directly by estimates of the slope coefficients.

The linear regression model specifies an additive (often specified to be independent and identically distributed) error so that, for the typical ith observation,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \ u_i, \quad i = 1, \ldots, N$$

The OLS estimator minimizes the sum of squared errors, $\sum_{i=1}^{N}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$.

Matrix notation provides a compact way to represent the estimator and variance matrix formulas that involve sums of products and cross products. We define the $N \times 1$ column vector $\mathbf{y}$ to have the ith entry y_i, and we define the $N \times K$ regressor matrix $\mathbf{X}$ to have the ith row $\mathbf{x}_i'$. Then the OLS estimator can be written in several ways, with

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \left(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{N}\mathbf{x}_i y_i \\
&= \begin{bmatrix}
\sum_{i=1}^{N} x_{1i}^2 & \sum_{i=1}^{N} x_{1i}x_{2i} & \cdots & \sum_{i=1}^{N} x_{1i}x_{Ki} \\
\sum_{i=1}^{N} x_{2i}x_{1i} & \sum_{i=1}^{N} x_{2i}^2 & & \vdots \\
\vdots & & \ddots & \vdots \\
\sum_{i=1}^{N} x_{Ki}x_{1i} & \cdots & \cdots & \sum_{i=1}^{N} x_{Ki}^2
\end{bmatrix}^{-1}
\begin{bmatrix}
\sum_{i=1}^{N} x_{1i}y_i \\
\sum_{i=1}^{N} x_{2i}y_i \\
\vdots \\
\sum_{i=1}^{N} x_{Ki}y_i
\end{bmatrix}
\end{aligned}$$

We define all vectors as column vectors, with a transpose if row vectors are desired. By contrast, Stata commands and Mata commands define vectors as row vectors, so in parts of Stata and Mata code, we need to take a transpose to conform to the notation in the book.

3.4.3 Properties of the OLS estimator

The properties of any estimator vary with the assumptions made about the DGP. For the linear regression model, this reduces to assumptions about the regression error u_i.

The starting point for analysis is to assume that u_i satisfies the following classical conditions:

1. $E(u_i|\mathbf{x}_i) = 0$ (exogeneity of regressors)
2. $E(u_i^2|\mathbf{x}_i) = \sigma^2$ (conditional homoskedasticity)
3. $E(u_i u_j|\mathbf{x}_i, \mathbf{x}_j) = 0$, $i \neq j$ (conditionally uncorrelated observations)

Assumption 1 is essential for consistent estimation of $\boldsymbol{\beta}$ and implies that the conditional mean given in (3.1) is correctly specified. This means that the conditional mean is linear and that all relevant variables have been included in the regression. Assumption 1 is relaxed in chapter 7. Assumptions 2 and 3 determine the form of the VCE of $\widehat{\boldsymbol{\beta}}$.

3.4.4 Default standard errors

Assumptions 1–3 lead to $\widehat{\boldsymbol{\beta}}$ being asymptotically normally distributed with the default estimator of the VCE

$$\widehat{V}_{\text{default}}\left(\widehat{\boldsymbol{\beta}}\right) = s^2(\mathbf{X}'\mathbf{X})^{-1}$$

where

$$s^2 = (N-K)^{-1}\sum_{i=1}^{N}\widehat{u}_i^2 \tag{3.2}$$

and $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$. Under assumptions 1–3, the OLS estimator is fully efficient. If, additionally, u_i is normally distributed, then "t statistics" are exactly t distributed. This fourth assumption is not made, but it is common to continue to use the t distribution in the hope that it provides a better approximation than the standard normal in finite samples.

When assumptions 2 and 3 are relaxed, OLS is no longer fully efficient. In chapter 6, we present examples of more efficient, feasible generalized least-squares estimation. In the current chapter, we continue to use the OLS estimator, as is often done in practice, but we use alternative estimates of the VCE that are valid when assumption 2, assumption 3, or both are relaxed, provided the sample size is sufficiently large for the relevant asymptotic theory to provide a good approximation.

3.4.5 Heteroskedasticity-robust standard errors

Given assumptions 1 and 3, but not 2, we have heteroskedastic uncorrelated errors. Then a robust estimator, or more precisely a heteroskedasticity-robust estimator, of the VCE of the OLS estimator is

$$\widehat{V}_{\text{robust}}\left(\widehat{\boldsymbol{\beta}}\right) = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{N}{N-K}\sum_{i=1}^{N}\widehat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1} \tag{3.3}$$

For cross-sectional data that are independent, this estimator, introduced by White (1980), has supplanted the default variance matrix estimate in most applied work because heteroskedasticity is the norm, and in that case, the default estimate of the VCE is incorrect.

In Stata, a robust estimate of the VCE is obtained by using the `vce(robust)` option of the `regress` command, as illustrated in section 3.5.2. Related options are `vce(hc2)` and `vce(hc3)`, which may provide better heteroskedasticity-robust estimates of the VCE when the sample size is small; see [R] **regress**. The robust estimator of the VCE has been extended to other estimators and models, and a feature of Stata is the `vce(robust)` option, which is applicable for many estimation commands. Some community-contributed commands use `robust` in place of `vce(robust)`.

3.4.6 Cluster–robust standard errors

When errors for different observations are correlated, assumption 3 is violated. Then both default and heteroskedastic robust estimates of the VCE are invalid, and different ways in which error correlation may arise lead to different robust estimates of the VCE. Various robust estimates of the VCE are presented in section 13.4.

For cross-sectional data, the most common violation of assumption 3 is that errors are clustered. Clustered or grouped errors are errors that are correlated within a cluster or group and are uncorrelated across clusters. A simple example of clustering arises when sampling is of independent units but errors for individuals within the unit are correlated. For example, 100 independent villages may be sampled, with several people from each village surveyed. Then, if a regression model overpredicts y for one village member, it is likely to overpredict for other members of the same village, indicating positive correlation. Similar comments apply when sampling is of households with several individuals in each household. Another leading example is panel data with independence over individuals but with correlation over time for a given individual.

Given assumption 1, but not 2 or 3, a cluster–robust estimator of the VCE of the OLS estimator is

$$\widehat{V}_{\text{cluster}}\left(\widehat{\boldsymbol{\beta}}\right) = (\mathbf{X}'\mathbf{X})^{-1}\left(\frac{G}{G-1}\frac{N-1}{N-K}\sum_{g=1}^{G}\mathbf{X}_g\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g'\right)(\mathbf{X}'\mathbf{X})^{-1}$$

where $g = 1, \ldots, G$ denotes the cluster (such as village), $\widehat{\mathbf{u}}_g$ is the vector of residuals for the observations in the gth cluster, and $\mathbf{X}_g$ is a matrix of the regressors for the

observations in the gth cluster. The key assumptions made are error independence across clusters and that the number of clusters $G \to \infty$.

Cluster–robust standard errors can be computed by using the `vce(cluster` *clustvar*`)` option in Stata, where clusters are defined by the different values taken by the *clustvar* variable. The estimate of the VCE is in fact heteroskedasticity-robust and cluster–robust because there is no restriction on $\text{Cov}(u_{gi}, u_{gj})$. The cluster VCE estimate can be applied to many estimators and models; see section 13.4.6.

Cluster–robust standard errors must be used when data are clustered. For a scalar regressor x, a rule of thumb is that cluster–robust standard errors are

$$\tau \simeq \sqrt{1 + \rho_x \rho_u (M-1)} \tag{3.4}$$

times the incorrect default standard errors, where ρ_x is the within-cluster correlation coefficient of the regressor, ρ_u is the within-cluster correlation coefficient of the error, and M is the average cluster size. This rule of thumb is a good guide in most settings, but when x is an experimentally assigned treatment with values that vary across observations within the same cluster, one should use the more general rule of thumb that $\tau \simeq \sqrt{1 + \rho_{xu}(M-1)}$, where ρ_{xu} is the within-cluster correlation of $x_i u_i$. Cluster–robust standard errors can be much larger than default or heteroskedastic–robust standard errors.

It can be necessary to use cluster–robust standard errors even where it is not immediately obvious. This is particularly the case when a regressor is an aggregated or macrovariable because then $\rho_x = 1$. For example, suppose we use data from the U.S. Current Population Survey and regress individual earnings on individual characteristics and a state-level regressor that does not vary within a state. Then, if there are many individuals in each state so M is large, even slight error correlation for individuals in the same state can lead to great downward bias in default standard errors and in heteroskedasticity-robust standard errors. Clustering can also be induced by the design of sample surveys. This topic is pursued in section 6.9.

Statistical inference for OLS based on cluster–robust standard errors uses critical values and p-values based on the t distribution with $(G-1)$ degrees of freedom, where G is the number of clusters. When there are few clusters, this approximation can lead to considerable underestimation of standard errors and associated test p-values and to confidence intervals that are too narrow. Better inference for OLS with few clusters is pursued in section 6.4.6 and in section 12.6. In particular, see section 12.6 for the community-contributed `boottest` command (Roodman et al. 2019), which implements a wild cluster bootstrap that can lead to better finite cluster inference.

Many microeconometric applications use clustered data. Then other estimators than OLS are often used, most notably fixed-effects and random-effects estimators. For linear models, these methods are presented in sections 6.5–6.7 and, for panel data, in chapter 8. For nonlinear models, see section 13.9 and, for panel data, see chapter 22. For the recently proposed design-based approach to inference, see section 24.4.7.

3.4.7 Bootstrap standard errors

An appropriate alternative way to compute heteroskedasticity-robust or cluster–robust standard errors is to use an appropriate bootstrap. This is a widely applicable method for obtaining standard errors and confidence intervals for parameters in cases where the asymptotic distribution is either not available or is available but is inconvenient to implement.

Here we present simple bootstraps that yield standard errors that are asymptotically equivalent to those obtained using the `vce(robust)` and `vce(cluster` *clustvar*`)` options. A refined bootstrap procedure, if feasible, provides an improvement over the usual asymptotic distribution. These distinctions are further developed and used in section 12.5.

The basic idea of the bootstrap is that the sample is used as a population, and we then obtain a number of samples from this "population" by repeatedly resampling observations with replacement. Such samples are referred to as bootstrap samples. This is a substitute for the ideal but impractical situation of having multiple independent samples. We then obtain the sampling distribution of the parameters of interest by fitting the same model to the many bootstrap samples. Moments of the distribution can then be computed from the collection of estimates.

Resampling from a given sample is easiest to understand in the independent and identically distributed setting with sample $y_i, i = 1, \ldots, N$. Suppose that the target parameter is the population mean, denoted μ, and the estimator $\widehat{\mu}$ is the sample mean $\overline{y}$. Then we can draw B different samples of N observations each by sampling with replacement. Each sample generates a sample mean, $\overline{y}_b$, $b = 1, \ldots, B$, so we have B independent estimates. Moments of the distribution of $\widehat{\mu}$ can then be computed given the empirical distribution of these B estimates.

Now consider the linear regression setting with data $(y_i, \mathbf{x}_i), i = 1, \ldots, N$, and model errors that are independent but heteroskedastic. A bootstrap called a paired bootstrap or nonparametric bootstrap obtains bootstrap resamples by sampling $(y_i, \mathbf{x}_i)$, jointly and with replacement. Each bootstrap sample of N observations generates an estimate of the regression parameters, denoted $\widehat{\boldsymbol{\beta}}_b$, $b = 1, \ldots, B$.

The bootstrap estimate of variance of an estimator is the usual formula for estimating a variance of (say) β_j, applied to the B bootstrap replications

$$s^2_{\widehat{\beta}_j} = \frac{1}{B-1} \sum_b \left(\widehat{\beta}_{jb} - \overline{\widehat{\beta}}_{jb} \right)^2$$

The bootstrap $100(1-\alpha)$ percent confidence interval for β_j is obtained by using the asymptotic α percent critical values from the standard normal distribution,

$$\left(\widehat{\beta}_j - z_{\alpha/2} \times s_{\widehat{\beta}_j}, \widehat{\beta}_j + z_{\alpha/2} \times s_{\widehat{\beta}_j} \right)$$

where $\Pr(Z > z_{\alpha/2}) = \alpha/2$. This bootstrap yields standard errors and confidence intervals that are asymptotically equivalent to those obtained using heteroskedastic–robust standard errors.

When regression model errors are instead clustered, the preceding method is adapted by resampling entire clusters with replacement. Each bootstrap sample of G clusters generates an estimate of the regression parameters, denoted $\widehat{\boldsymbol{\beta}}_b$, $b = 1, \ldots, B$. Then $s^2_{\widehat{\beta}_j}$ and the associated confidence interval are computed using the preceding formulas. This cluster pairs bootstrap yields standard errors and confidence intervals that are equivalent as $G \to \infty$ to those obtained using cluster–robust standard errors.

3.4.8 Regression in logs

The medical expenditure data are very right skewed. Then a linear model in levels can provide very poor predictions because it restricts the effects of regressors to be additive. For example, aging 10 years is assumed to increase medical expenditures by the same amount regardless of observed health status. Instead, it is more reasonable to assume that aging 10 years has a multiplicative effect. For example, it may increase medical expenditures by 20%.

We begin with an exponential mean model for positive expenditures, with error that is also multiplicative, so $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})\varepsilon_i$. Defining $\varepsilon_i = \exp(u_i)$, we have $y_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + u_i)$, and taking the natural logarithm, we fit the log-linear model

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

by OLS regression of $\ln y$ on $\mathbf{x}$. The conditional mean of $\ln y$ is being modeled, rather than the conditional mean of y. In particular,

$$E(\ln y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$$

assuming u_i has conditional mean zero.

Parameter interpretation requires care. For regression of $\ln y$ on $\mathbf{x}$, the coefficient β_j measures the effect of a change in regressor x_j on $E(\ln y|\mathbf{x})$, but ultimate interest lies instead on the effect on $E(y|\mathbf{x})$. Some algebra shows that β_j measures the proportionate change in $E(y|\mathbf{x})$ as x_j changes, called a semielasticity, rather than the level of change in $E(y|\mathbf{x})$. For example, if $\beta_j = 0.02$, then a one-unit change in x_j is associated with a proportionate increase of 0.02, or a 2% increase, in $E(y|\mathbf{x})$.

Prediction of $E(y|\mathbf{x})$ is substantially more difficult because it can be shown that $E(\ln y|\mathbf{x}) \neq \exp(\mathbf{x}'\boldsymbol{\beta})$. This is pursued in section 4.2.3. Buntin and Zaslavsky (2004) compare several alternative regression models for medical expenditures.

3.5 Basic regression analysis

We use `regress` to run an OLS regression of the natural logarithm of medical expenditures, `ltotexp`, on `suppins` and several demographic and health-status measures. Using