Sources:

1) CT MUS (Cameron and Trivedi, Microeconometrics Using Stata)

2) CT (Cameron and Trivedi, Microeconometrics)

3) Wooldridge-Introductory Econometrics

Reading data into Stata and examining data

From CT MUS (Cameron and Trivedi, Microeconometrics Using Stata)

```
. * Variable description for medical expenditure dataset
. use mus03data.dta
. describe totexp ltotexp posexp suppins phylim actlim totchr age female income
```

desc

gen ltot =ln(totexp)

summarize (can be abbreviated as 'sum'; for example: 'sum posexp')

summarize p*

list if in 1/10, clean

list suppins phylim

tabulate (can be abbreviated as 'tab': for example: 'tab posexp')

tab income if income<=0

sum totexp, detail

table female totchr

table female totchr suppins

tabulate female suppins, row col

table female, contents(N totchr mean totchr sd totchr p50 totchr)

table female suppins, contents(N totchr mean totchr sd totchr p50 totchr)

tabstat totexp ltotexp, stat (count mean)

tabstat totexp ltotexp, stat (count mean) col(stat)

Notation and Conventions (from CT)

1.6. Notation and Conventions

Vector and matrix algebra are used extensively.

Vectors are defined as column vectors and represented using lowercase bold. For example, for linear regression the regressor vector $\mathbf{x}$ is a $K \times 1$ column vector with jth entry x_j and the parameter vector $\boldsymbol{\beta}$ is a $K \times 1$ column vector with jth entry β_j, so

$$\underset{(K \times 1)}{\mathbf{x}} = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \quad \text{and} \quad \underset{(K \times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}.$$

Then the linear regression model $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$ is expressed as $y = \mathbf{x}'\boldsymbol{\beta} + u$. At times a subscript i is added to denote the typical ith observation. The linear regression equation for the ith observation is then

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

The sample is one of N observations, $\{(y_i, \mathbf{x}_i), i = 1, \ldots, N\}$. In this book observations are usually assumed to be independent over i.

Matrices are represented using uppercase bold. In matrix notation the sample is $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ is an $N \times 1$ vector with ith entry y_i and $\mathbf{X}$ is a matrix with ith row $\mathbf{x}_i'$, so

$$\underset{(N \times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \text{and} \quad \underset{(N \times \dim(\mathbf{x}))}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix}.$$

The linear regression model upon stacking all N observations is then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where $\mathbf{u}$ is an $N \times 1$ column vector with ith entry u_i.

Matrix notation is compact but at times it is clearer to write products of matrices as summations of products of vectors. For example, the OLS estimator can be equivalently written in either of the following ways:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i.$$

Generic notation for a parameter is the $q \times 1$ vector $\boldsymbol{\theta}$. The regression parameters are represented by the $K \times 1$ vector $\boldsymbol{\beta}$, which may equal $\boldsymbol{\theta}$ or may be a subset of $\boldsymbol{\theta}$ depending on the context.

3.3.1 Basic regression theory

We begin by introducing terminology used throughout the rest of this book. Let θ denote the vector of parameters to be estimated, and let $\widehat{\theta}$ denote an estimator of θ. Ideally, the distribution of $\widehat{\theta}$ is centered on θ with small variance, for precision, and a known distribution, to permit statistical inference. We restrict analysis to estimators that are consistent for θ, meaning that in infinitely large samples, $\widehat{\theta}$ equals θ aside from negligible random variation. This is denoted by $\widehat{\theta} \xrightarrow{p} \theta$ or more formally by $\widehat{\theta} \xrightarrow{p} \theta_0$, where θ_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, 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

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

where $\text{Var}(\widehat{\theta})$ 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 $\widehat{V}(\widehat{\theta})$. Standard errors of the parameter estimates are obtained as the square root of diagonal entries in $\widehat{V}(\widehat{\theta})$. Different assumptions about the data-generating process (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 it may provide a better approximation in smaller samples.

3.3.2 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 marginal effect 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 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

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

(Vectors are defined as column vectors here.)

3.3.3 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 6.

Assumptions 2 and 3 determine the form of the VCE of $\widehat{\boldsymbol{\beta}}$. Assumptions 1–3 lead to $\widehat{\boldsymbol{\beta}}$ being asymptotically normally distributed with the default estimator of the VCE

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

where

$$s^2 = (N-k)^{-1}\sum_i \widehat{u}_i^2 \tag{3.2}$$

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 5, we present examples of more-efficient feasible generalized least-squares (FGLS) 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.

3.3.4 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

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

For cross-section 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.

3.3.5 Cluster–robust standard errors

When errors for different observations are correlated, assumption 3 is violated. Then both default and robust estimates of the VCE are invalid. For time-series data, this is the case if errors are serially correlated, and the newey command should be used. For cross-section data, this can arise when 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

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

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 9.6.

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 $\sqrt{1+\rho_x\rho_u(M-1)}$ 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.

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 macro variable, 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 5.5.

3.3.6 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 is independent with 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 2%, in $E(y|\mathbf{x})$.

3.4.2 The regress command

The `regress` command performs OLS regression and yields an analysis-of-variance table, goodness-of-fit statistics, coefficient estimates, standard errors, t statistics, p-values, and confidence intervals. The syntax of the command is

`regress` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

Other Stata estimation commands have similar syntaxes. The output from `regress` is similar to that from many linear regression packages.

********** **Run the following regression in Stata:** **********

```
reg ltotexp suppins phylim actlim totchr age female income, robust
```

The regressors are jointly statistically significant, because the overall F statistic of 126.97 has a p-value of 0.000. At the same time, much of the variation is unexplained with $R^2 = 0.2289$. The root MSE statistic reports s, the standard error of the regression, defined in (3.2). By using a two-sided test at level 0.05, all regressors are individually statistically significant because $p < 0.05$, aside from age and `female`. The strong statistical insignificance of age may be due to sample restriction to elderly people and the inclusion of several health-status measures that capture well the health effect of age.

Statistical significance of coefficients is easily established. More important is the economic significance of coefficients, meaning the measured impact of regressors on medical expenditures. This is straightforward for regression in levels, because we can directly use the estimated coefficients. But here the regression is in logs. From section 3.3.6, in the log-linear model, parameters need to be interpreted as semielasticities. For example, the coefficient on suppins is 0.256. This means that private supplementary insurance is associated with a 0.256 proportionate rise, or a 25.6% rise, in medical expenditures. Similarly, large effects are obtained for the health-status measures, whereas health expenditures for women are 8.4% lower than those for men after controlling for other characteristics. The income coefficient of 0.0025 suggests a very small effect, but this is misleading. The standard deviation of income is 22, so a 1-standard deviation in income leads to a 0.055 proportionate rise, or 5.5% rise, in medical expenditures.

MEs in nonlinear models are discussed in more detail in section 10.6. The preceding interpretations are based on calculus methods that consider very small changes in the regressor. For larger changes in the regressor, the finite-difference method is more appropriate. Then the interpretation in the log-linear model is similar to that for the exponential conditional mean model; see section 10.6.4. For example, the estimated effect of going from no supplementary insurance (suppins=0) to having supplementary insurance (suppins=1) is more precisely a $100 \times (e^{0.256} - 1)$, or 29.2%, rise.

The regress command provides additional results that are not listed. In particular, the estimate of the VCE is stored in the matrix e(V). Ways to access this and other stored results from regression have been given in section 1.6. Various postestimation commands enable prediction, computation of residuals, hypothesis testing, and model specification tests. Many of these are illustrated in subsequent sections. Two useful commands are

```
. * Display stored results and list available postestimation commands
. ereturn list
  (output omitted)
. help regress postestimation
  (output omitted)
```

3.4.3 Hypothesis tests

The test command performs hypothesis tests using the Wald test procedure that uses the estimated model coefficients and VCE. We present some leading examples here, with a more extensive discussion deferred to section 12.3. The F statistic version of the Wald test is used after regress, whereas for many other estimators the chi-squared version is instead used.

A common test is one of equality of coefficients. For example, consider testing that having a functional limitation has the same impact on medical expenditures as having an activity limitation. The test of $H_0: \beta_{\texttt{phylim}} = \beta_{\texttt{actlim}}$ against $H_a: \beta_{\texttt{phylim}} \neq \beta_{\texttt{actlim}}$ is implemented as

```
test phylim = actlim
```

Because $p = 0.61 > 0.05$, we do not reject the null hypothesis at the 5% significance level. There is no statistically significant difference between the coefficients of the two variables.

Another common test is one of the joint statistical significance of a subset of the regressors. A test of the joint significance of the health-status measures is one of H_0: $\beta_{\texttt{phylim}} = 0$, $\beta_{\texttt{actlim}} = 0$, $\beta_{\texttt{totchr}} = 0$ against H_a: at least one is nonzero. This is implemented as

```
. *  Joint test of statistical significance of several variables
. test phylim actlim totchr
 ( 1)  phylim = 0
 ( 2)  actlim = 0
 ( 3)  totchr = 0
       F(  3,  2947) =  272.36
            Prob > F =    0.0000
```

These three variables are jointly statistically significant at the 0.05 level because $p = 0.000 < 0.05$.

3.4.4 Tables of output from several regressions

It is very useful to be able to tabulate key results from multiple regressions for both one's own analysis and final report writing.

The `estimates store` command after regression leads to results in `e()` being associated with a user-provided model name and preserved even if subsequent models are fitted. Given one or more such sets of stored estimates, `estimates table` presents a table of regression coefficients (the default) and, optionally, additional results. The `estimates stats` command lists the sample size and several likelihood-based statistics.

We compare the original regression model with a variant that replaces `income` with `educyr`. The example uses several of the available options for `estimates table`.

```
. * Store and then tabulate results from multiple regressions
. quietly regress ltotexp suppins phylim actlim totchr age female income,
> vce(robust)
. estimates store REG1
. quietly regress ltotexp suppins phylim actlim totchr age female educyr,
> vce(robust)
. estimates store REG2
. estimates table REG1 REG2, b(%9.4f) se stats(N r2 F ll)
> keep(suppins income educyr)
```

Variable	REG1	REG2
suppins	0.2556	0.2063
	0.0466	0.0471
income	0.0025	
	0.0010	
educyr		0.0480
		0.0070
N	2955.0000	2955.0000
r2	0.2289	0.2406
F	126.9723	132.5337
ll	-4.73e+03	-4.71e+03

legend: b/se

3.6 Prediction

For the linear regression model, the estimator of the conditional mean of y given $\mathbf{x} = \mathbf{x}_p$, $E(y|\mathbf{x}_p) = \mathbf{x}_p'\beta$, is the conditional predictor $\widehat{y} = \mathbf{x}_p'\widehat{\beta}$. We focus here on prediction for each observation in the sample. We begin with prediction from a linear model for medical expenditures, because this is straightforward, before turning to the log-linear model.

3.6.1 In-sample prediction

The most common type of prediction is in-sample, where evaluation is at the observed regressor values for each observation. Then $\widehat{y}_i = \mathbf{x}_i'\widehat{\beta}$ predicts $E(y_i|\mathbf{x}_i)$ for $i = 1, \ldots, N$.

To do this, we use `predict` after `regress`. The syntax for `predict` is

`predict` [*type*] *newvar* [*if*] [*in*] [, *options*]

The user always provides a name for the created variable, *newvar*. The default option is the prediction $\widehat{y}_i$. Other options yield residuals (usual, standardized, and studentized), several leverage and influential observation measures, predicted values, and associated standard errors of prediction. We have already used some of these options in section 3.5. The `predict` command can also be used for out-of-sample prediction. When used for in-sample prediction, it is good practice to add the `if e(sample)` qualifier, because this ensures that prediction is for the same sample as that used in estimation.

We consider prediction based on a linear regression model in levels rather than logs. We begin by reporting the regression results with `totexp` as the dependent variable.

```
. * Change dependent variable to level of positive medical expenditures
. use mus03data.dta, clear
. keep if totexp > 0
(109 observations deleted)
```

```
. regress totexp suppins phylim actlim totchr age female income, vce(robust)
Linear regression                               Number of obs =    2955
                                                F(  7,  2947) =   40.58
                                                Prob > F      =  0.0000
                                                R-squared     =  0.1163
                                                Root MSE      =   11285
```

totexp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
suppins	724.8632	427.3045	1.70	0.090	-112.9824	1562.709
phylim	2389.019	544.3493	4.39	0.000	1321.675	3456.362
actlim	3900.491	705.2244	5.53	0.000	2517.708	5283.273
totchr	1844.377	186.8938	9.87	0.000	1477.921	2210.832
age	-85.36264	37.81868	-2.26	0.024	-159.5163	-11.20892
female	-1383.29	432.4759	-3.20	0.001	-2231.275	-535.3044
income	6.46894	8.570658	0.75	0.450	-10.33614	23.27402
_cons	8358.954	2847.802	2.94	0.003	2775.07	13942.84

We then predict the level of medical expenditures:

```
. * Prediction in model linear in levels
. predict yhatlevels
(option xb assumed; fitted values)
. summarize totexp yhatlevels
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	2955	7290.235	11990.84	3	125610
yhatlevels	2955	7290.235	4089.624	-236.3781	22559

The summary statistics show that on average the predicted value `yhatlevels` equals the dependent variable. This suggests that the predictor does a good job. But this is misleading because this is always the case after OLS regression in a model with an intercept, since then residuals sum to zero implying $\sum y_i = \sum \widehat{y}_i$. The standard deviation of `yhatlevels` is \$4,090, so there is some variation in the predicted values.

For this example, a more discriminating test is to compare the median predicted and actual values. We have

```
. * Compare median prediction and median actual value
. tabstat totexp yhatlevels, stat (count p50) col(stat)
```

variable	N	p50
totexp	2955	3334
yhatlevels	2955	6464.692

There is considerable difference between the two, a consequence of the right-skewness of the original data, which the linear regression model does not capture.

The `stdp` option provides the standard error of the prediction, and the `stdf` option provides the standard error of the prediction for each sample observation, provided the

original estimation command used the default VCE. We therefore reestimate without vce(robust) and use predict to obtain

```
. * Compute standard errors of prediction and forecast with default VCE
. quietly regress totexp suppins phylim actlim totchr age female income
. predict yhatstdp, stdp
. predict yhatstdf, stdf
. summarize yhatstdp yhatstdf
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhatstdp	2955	572.7	129.6575	393.5964	2813.983
yhatstdf	2955	11300.52	10.50946	11292.12	11630.8

The first quantity views $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ as an estimate of the conditional mean $\mathbf{x}_i'\boldsymbol{\beta}$ and is quite precisely estimated because the average standard deviation is \$573 compared with an average prediction of \$7,290. The second quantity views $\mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ as an estimate of the actual value y_i and is very imprecisely estimated because $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$, and the error u_i here has relatively large variance since the levels equation has $s = 11285$.

More generally, microeconometric models predict poorly for a given individual, as evidenced by the typically low values of R^2 obtained from regression on cross-section data. These same models may nonetheless predict the conditional mean well, and it is this latter quantity that is needed for policy analysis that focuses on average behavior.

3.6.2 Marginal effects

The mfx postestimation command calculates MEs and elasticities evaluated at sample means, along with associated standard errors and confidence intervals where relevant. The default is to obtain these for the quantity that is the default for predict. For many estimation commands, including regress, this is the conditional mean. Then mfx computes for each continuous regressor $\partial E(y|\mathbf{x})/\partial x$, and for 0/1 indicator variables $\Delta E(y|\mathbf{x})$, evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and $\mathbf{x} = \overline{\mathbf{x}}$.

For the linear model, the estimated ME of the jth regressor is $\widehat{\beta}_j$, so there is no need to use mfx. But mfx can also be used to compute elasticities and semielasticities. For example, the eyex option computes the elasticity $\partial y/\partial x \times (x/y)$, evaluated at sample means, which equals $\widehat{\beta}_j \times (\overline{x}_j/\overline{y})$ for the linear model. We have

```
. * Compute elasticity for a specified regressor
. quietly regress totexp suppins phylim actlim totchr age female income,
> vce(robust)
. mfx, varlist(totchr) eyex
Elasticities after regress
      y  = Fitted values (predict)
         =  7290.2352
```

variable	ey/ex	Std. Err.	z	P>|z|	[95% C.I.]		X
totchr	.457613	.04481	10.21	0.000	.369793	.545433	1.8088

A 1% increase in chronic problems is associated with a 0.46% increase in medical expenditures. The `varlist(totchr)` option restricts results to just the regressor `totchr`.

The `predict()` option of `mfx` allows the computation of MEs for the other quantities that can be produced using `predict`.

3.6.3 Prediction in logs: The retransformation problem

Transforming the dependent variable by taking the natural logarithm complicates prediction. It is easy to predict $E(\ln y|\mathbf{x})$, but we are instead interested in $E(y|\mathbf{x})$ because we want to predict the level of medical expenditures rather than the natural logarithm. The obvious procedure of predicting $\ln y$ and taking the exponential is wrong because $\exp\{E(\ln y)\} \neq E(y)$, just as, for example, $\sqrt{E(y^2)} \neq E(y)$.

The log-linear model $\ln y = \mathbf{x}'\boldsymbol{\beta} + u$ implies that $y = \exp(\mathbf{x}'\boldsymbol{\beta})\exp(u)$. It follows that

$$E(y_i|\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta})E\{\exp(u_i)\}$$

The simplest prediction is $\exp(\mathbf{x}_i'\widehat{\boldsymbol{\beta}})$, but this is wrong because it ignores the multiple $E\{\exp(u_i)\}$. If it is assumed that $u_i \sim N(0,\sigma^2)$, then it can be shown that $E\{\exp(u_i)\} = \exp(0.5\sigma^2)$, which can be estimated by $\exp(0.5\widehat{\sigma}^2)$, where $\widehat{\sigma}^2$ is an unbiased estimator of the log-linear regression model error. A weaker assumption is to assume that u_i is independent and identically distributed, in which case we can consistently estimate $E\{\exp(u_i)\}$ by the sample average $N^{-1}\sum_{j=1}^{N}\exp(\widehat{u}_j)$; see Duan (1983).

Applying these methods to the medical expenditure data yields

```
. * Prediction in levels from a logarithmic model
. quietly regress ltotexp suppins phylim actlim totchr age female income
. quietly predict lyhat
. generate yhatwrong = exp(lyhat)
. generate yhatnormal = exp(lyhat)*exp(0.5*e(rmse)^2)
. quietly predict uhat, residual
. generate expuhat = exp(uhat)
. quietly summarize expuhat
. generate yhatduan = r(mean)*exp(lyhat)
. summarize totexp yhatwrong yhatnormal yhatduan yhatlevels
```

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	2955	7290.235	11990.84	3	125610
yhatwrong	2955	4004.453	3303.555	959.5991	37726.22
yhatnormal	2955	8249.927	6805.945	1976.955	77723.13
yhatduan	2955	8005.522	6604.318	1918.387	75420.57
yhatlevels	2955	7290.235	4089.624	-236.3781	22559

Ignoring the retransformation bias leads to a very poor prediction, because yhatwrong has a mean of $4,004 compared with the sample mean of $7,290. The two alternative methods yield much closer average values of $8,250 and $8,006. Furthermore, the predictions from log regression, compared with those in levels, have the desirable feature of always being positive and have greater variability. The standard deviation of yhatnormal, for example, is $6,806 compared with $4,090 from the levels model.

3.6.4 Prediction exercise

There are several ways that predictions can be used to simulate the effects of a policy experiment. We consider the effect of a binary treatment, whether a person has supplementary insurance, on medical expenditure. Here we base our predictions on estimates that assume supplementary insurance is exogenous. A more thorough analysis could instead use methods that more realistically permit insurance to be endogenous. As we discuss in section 6.2.1, a variable is endogenous if it is related to the error term. Our analysis here assumes that supplementary insurance is not related to the error term.

An obvious comparison is to compare the difference in sample means $(\overline{y}_1 - \overline{y}_0)$, where the subscript 1 denotes those with supplementary insurance and the subscript 0 denotes those without supplementary insurance. This measure does not control for individual characteristics. A measure that does control for individual characteristics is the difference in mean predictions $(\widehat{\overline{y}}_1 - \widehat{\overline{y}}_0)$, where, for example, $\widehat{\overline{y}}_1$ denotes the average prediction for those with health insurance.

We implement the first two approaches for the complete sample based on OLS regression in levels and in logs. We obtain

```
. * Predicted effect of supplementary insurance: methods 1 and 2
. bysort suppins: summarize totexp yhatlevels yhatduan
```

-> suppins = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	1207	6824.303	11425.94	9	104823
yhatlevels	1207	6824.303	4077.064	-236.3781	20131.43
yhatduan	1207	6745.959	5365.255	1918.387	54981.73

-> suppins = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
totexp	1748	7611.963	12358.83	3	125610
yhatlevels	1748	7611.963	4068.397	502.9237	22559
yhatduan	1748	8875.255	7212.993	2518.538	75420.57

The average difference is $788 (from 7612 – 6824) using either the difference in sample means or the difference in fitted values from the linear model. Equality of the two is a consequence of OLS regression and prediction using the estimation sample. The log-linear model, using the prediction based on Duan's method, gives a larger average difference of $2,129 (from 8875 – 6746).

A third measure is the difference between the mean predictions, one with suppins set to 1 for all observations and one with suppins = 0. For the linear model, this is simply the estimated coefficient of suppins, which is $725.

For the log-linear model, we need to make separate predictions for each individual with suppins set to 1 and with suppins set to 0. For simplicity, we make predictions in levels from the log-linear model assuming normally distributed errors. To make these changes and after the analysis have suppins returned to its original sample values, we use preserve and restore (see section 2.5.2). We obtain

```
. * Predicted effect of supplementary insurance: method 3 for log-linear model
. quietly regress ltotexp suppins phylim actlim totchr age female income
. preserve
. quietly replace suppins = 1
. quietly predict lyhat1
. generate yhatnormal1 = exp(lyhat1)*exp(0.5*e(rmse)^2)
. quietly replace suppins = 0
. quietly predict lyhat0
. generate yhatnormal0 = exp(lyhat0)*exp(0.5*e(rmse)^2)
. generate treateffect = yhatnormal1 - yhatnormal0
. summarize yhatnormal1 yhatnormal0 treateffect
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhatnormal1	2955	9077.072	7313.963	2552.825	77723.13
yhatnormal0	2955	7029.453	5664.069	1976.955	60190.23
treateffect	2955	2047.619	1649.894	575.8701	17532.91

```
. restore
```

While the average treatment effect of $2,048 is considerably larger than that obtained by using the difference in sample means of the linear model, it is comparable to the estimate produced by Duan's method.

Note: If a random variable $X \sim N(\mu, \sigma^2)$, then Y = exp(X) ~ lognormal. $E(Y) = exp\left(\mu + \frac{1}{2}\sigma^2\right)$.

Note: If you need to run a regression but do not need the output to be displayed, use:

quietly regress y x1 x2 .....

4.3. Example: Returns to Schooling

A leading linear regression application from labor economics concerns measuring the impact of education on wages or earnings.

A typical returns to schooling model specifies

$$\ln w_i = \alpha s_i + \mathbf{x}_{2i}'\boldsymbol{\beta} + u_i, \quad i = 1, ..., N, \tag{4.5}$$

where w denotes hourly wage or annual earnings, s denotes years of completed schooling, and $\mathbf{x}_2$ denotes control variables such as work experience, gender, and family background. The subscript i denotes the ith person in the sample. Since the dependent variable is log wage, the model is a log-linear model and the coefficient α measures the proportionate change in earnings associated with a one-year increase in education.

Estimation of this model is most often by ordinary least squares. The transformation to $\ln w$ in practice ensures that errors are approximately homoskedastic, but it is still best to obtain heteroskedastic consistent standard errors as detailed in Section 4.4. Estimation can also be by quantile regression (see Section 4.6), if interest lies in distributional issues such as behavior in the lower quartile.

The regression (4.5) can be used immediately in a descriptive manner. For example, if $\widehat{\alpha} = 0.10$ then a one-year increase in schooling is associated with 10% higher earnings, controlling for all the factors included in $\mathbf{x}_2$. It is important to add the last qualifier as in this example the estimate $\widehat{\alpha}$ usually becomes smaller as $\mathbf{x}_2$ is expanded to include additional controls likely to influence earnings.

Policy interest lies in determining the impact of an *exogenous change* in schooling on earnings. However, schooling is not randomly assigned; rather, it is an outcome that depends on choices made by the individual. Human capital theory treats schooling as investment by individuals in themselves, and α is interpreted as a measure of return to human capital. The regression (4.5) is then a regression of one endogenous variable, y, on another, s, and so does not measure the causal impact of an exogenous change in s. The conditional mean function here is not causally meaningful because one is conditioning on a factor, schooling, that is *endogenous*. Indeed, unless we can argue that s is itself a function of variables at least one of which can vary independently of u, it is unclear just what it means to regard α as a causal parameter.

Such concern about endogenous regressors with observational data on individuals pervades microeconometric analysis. The standard assumptions of the linear regression model given in Section 4.4 are that regressors are exogenous. The consequences of endogenous regressors are considered in Section 4.7. One method to control for endogenous regressors, instrumental variables, is detailed in Section 4.8. A recent extensive review of ways to control for endogeneity in this wage–schooling example is given in Angrist and Krueger (1999). These methods are summarized in Section 2.8 and presented throughout this book.

4.4.1. Linear Regression Model

In a standard cross-section regression model with N observations on a scalar dependent variable and several regressors, the data are specified as $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ denotes observations on the dependent variable and $\mathbf{X}$ denotes a matrix of explanatory variables.

The general regression model with additive errors is written in vector notation as

$$\mathbf{y} = \mathrm{E}[\mathbf{y}|\mathbf{X}] + \mathbf{u}, \tag{4.6}$$

where $\mathrm{E}[\mathbf{y}|\mathbf{X}]$ denotes the conditional expectation of the random variable $\mathbf{y}$ given $\mathbf{X}$, and $\mathbf{u}$ denotes a vector of unobserved random errors or disturbances. The right-hand side of this equation decomposes $\mathbf{y}$ into two components, one that is deterministic given the regressors and one that is attributed to random variation or noise. We think of $\mathrm{E}[\mathbf{y}|\mathbf{X}]$ as a conditional prediction function that yields the average value, or more formally the expected value, of $\mathbf{y}$ given $\mathbf{X}$.

A **linear regression model** is obtained when $\mathrm{E}[\mathbf{y}|\mathbf{X}]$ is specified to be a linear function of $\mathbf{X}$. Notation for this model has been presented in detail in Section 1.6. In vector notation the ith observation is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \tag{4.7}$$

where $\mathbf{x}_i$ is a $K \times 1$ **regressor vector** and $\boldsymbol{\beta}$ is a $K \times 1$ **parameter vector**. At times it is simpler to drop the subscript i and write the model for typical observation as $y = \mathbf{x}'\boldsymbol{\beta} + u$. In matrix notation the N observations are stacked by row to yield

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{4.8}$$

where $\mathbf{y}$ is an $N \times 1$ **vector of dependent variables**, $\mathbf{X}$ is an $N \times K$ **regression matrix**, and $\mathbf{u}$ is an $N \times 1$ **error vector**.

Equations (4.7) and (4.8) are equivalent expressions for the linear regression model and will be used interchangeably. The latter is more concise and is usually the most convenient representation.

In this setting y is referred to as the **dependent variable** or **endogenous variable** whose variation we wish to study in terms of variation in $\mathbf{x}$ and u; u is referred to as the **error term** or **disturbance term**; and $\mathbf{x}$ is referred to as **regressors** or **predictors** or **covariates**. If Assumption 4 in Section 4.4.6 holds, then all components of $\mathbf{x}$ are **exogenous variables** or **independent variables**.

4.4.2. OLS Estimator

The OLS estimator is defined to be the estimator that minimizes the sum of squared errors

$$\sum_{i=1}^{N} u_i^2 = \mathbf{u}'\mathbf{u} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{4.9}$$

Setting the derivative with respect to $\boldsymbol{\beta}$ equal to $\mathbf{0}$ and solving for $\boldsymbol{\beta}$ yields the OLS estimator,

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{4.10}$$

see Exercise 4.5 for a more general result, where it is assumed that the matrix inverse of $\mathbf{X}'\mathbf{X}$ exists. If $\mathbf{X}'\mathbf{X}$ is of less than full rank, the inverse can be replaced by a generalized inverse. Then OLS estimation still yields the optimal linear predictor of y given $\mathbf{x}$ if squared error loss is used, but many different linear combinations of $\mathbf{x}$ will yield this optimal predictor.

4.4.4. Distribution of the OLS Estimator

We focus on the asymptotic properties of the OLS estimator. Consistency is established and then the limit distribution is obtained by rescaling the OLS estimator. Statistical inference then requires consistent estimation of the variance matrix of the estimator. The analysis makes extensive use of asymptotic theory, which is summarized in Appendix A.

Consistency

The properties of an estimator depend on the process that actually generated the data, the **data generating process (dgp).** We assume the dgp is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, so that the model (4.8) is correctly specified. In some places, notably Chapters 5 and 6 and Appendix A the subscript 0 is added to $\boldsymbol{\beta}$, so the dgp is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}$. See Section 5.2.3 for discussion.

Then

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},\end{aligned}$$

and the OLS estimator can be expressed as

$$\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \tag{4.11}$$

To prove consistency we rewrite (4.11) as

$$\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \left(N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1} N^{-1}\mathbf{X}'\mathbf{u}. \tag{4.12}$$

The reason for renormalization in the right-hand side is that $N^{-1}\mathbf{X}'\mathbf{X} = N^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i'$ is an average that converges in probability to a finite nonzero matrix if $\mathbf{x}_i$ satisfies assumptions that permit a law of large numbers to be applied to $\mathbf{x}_i\mathbf{x}_i'$ (see Section 4.4.8 for detail). Then

$$\operatorname{plim}\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + \left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u}\right),$$

using Slutsky's Theorem (Theorem A.3). The OLS estimator is **consistent** for $\boldsymbol{\beta}$ (i.e., $\operatorname{plim}\widehat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta}$) if

$$\operatorname{plim} N^{-1}\mathbf{X}'\mathbf{u} = \mathbf{0}. \tag{4.13}$$

If a law of large numbers can be applied to the average $N^{-1}\mathbf{X}'\mathbf{u} = N^{-1}\sum_i \mathbf{x}_i u_i$ then a necessary condition for (4.13) to hold is that $\mathrm{E}[\mathbf{x}_i u_i] = \mathbf{0}$.

APPENDIX A

Asymptotic Theory

A.2. Convergence in Probability

Because of the intrinsic randomness of a sample we can never be certain that a sequence b_N, such as an estimator $\widehat{\theta}$ (often denoted $\widehat{\theta}_N$ to make clear that it is a sequence), will be within a given small distance of its limit, even if the sample is infinitely large. However, we can be almost certain. Different ways of expressing this near certainty correspond to different types of convergence of a sequence of random variables to a limit. The one most used in econometrics is convergence in probability.

A.2.1. Convergence in Probability

Recall that a sequence of nonstochastic real numbers $\{a_N\}$ converges to a if, for any $\varepsilon > 0$, there exists $N^* = N^*(\varepsilon)$ such that, for all $N > N^*$,

$$|a_N - a| < \varepsilon.$$

For example, if $a_N = 2 + 3/N$, then the limit is $a = 2$ since $|a_N - a| = |2 + 3/N - 2| = |3/N| < \varepsilon$ for all $N > N^* = 3/\varepsilon$.

When more generally we have a sequence of random variables we cannot be certain of being within ε of the limit, even for large N, because of intrinsic randomness. Instead, we require that the probability of being within ε is arbitrarily close to one. Thus we require

$$\lim_{N \to \infty} \Pr[|b_N - b| < \varepsilon] = 1,$$

for any $\varepsilon > 0$. A formal definition is the following:

Definition A.1 (Convergence in Probability): A sequence of random variables $\{b_N\}$ **converges in probability** to b if, for any $\varepsilon > 0$ and $\delta > 0$, there exists $N^* = N^*(\varepsilon, \delta)$ such that, for all $N > N^*$,

$$\Pr[|b_N - b| < \varepsilon] > 1 - \delta. \tag{A.1}$$

We write plim $b_N = b$, where plim is shorthand for **probability limit**, or $b_N \xrightarrow{p} b$.

Note that b may be a constant or a random variable. Convergence in probability includes as a special case the usual definition of convergence for a sequence of real variables.

Definition A.1 is for a sequence of *scalar* random variables. The extension to **vector random variables**, such as a parameter vector estimator, is straightforward. We can either apply the theory for each element of $\mathbf{b}_N$, or replace $|b_N - b|$ by the scalar $(\mathbf{b}_N - \mathbf{b})'(\mathbf{b}_N - \mathbf{b}) = (b_{1N} - b_1)^2 + \cdots + (b_{KN} - b_K)^2$ or its square root $\|\mathbf{b}_N - \mathbf{b}\|$.

When the sequence $\{\mathbf{b}_N\}$ is a sequence of parameter estimates $\widehat{\boldsymbol{\theta}}$, we have the following large sample analogue of unbiasedness.

Definition A.2 (Consistency): An estimator $\widehat{\boldsymbol{\theta}}$ is **consistent** for $\boldsymbol{\theta}_0$ if

$$\text{plim}\, \widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0. \tag{A.2}$$

The subscript 0 on $\boldsymbol{\theta}$ is explained in Section 5.2.3. Note that unbiasedness need not imply consistency. Unbiasedness states only that the expected value of $\widehat{\boldsymbol{\theta}}$ is $\boldsymbol{\theta}_0$, and it permits variability around $\boldsymbol{\theta}_0$ that need not disappear as the sample size goes to infinity. Also, a consistent estimator need not be unbiased. For example, adding $1/N$ to an unbiased and consistent estimator produces a new estimator that is biased but still consistent.

Although the sequence of vector random variables $\{\mathbf{b}_N\}$ may converge to a random variable $\mathbf{b}$, in many econometric applications $\{\mathbf{b}_N\}$ converges to a constant. For example, we hope that an estimator of a parameter will converge in probability to the parameter itself. One should be aware that some of the results that follow apply only if the limit value $\mathbf{b}$ is a constant.

Theorem A.3 (Slutsky's Theorem): *Let* $\mathbf{b}_N$ *be a finite-dimensional vector of random variables, and* $g(\cdot)$ *be a real-valued function continuous at a constant vector point* $\mathbf{b}$. *Then*

$$\mathbf{b}_N \xrightarrow{P} \mathbf{b} \Rightarrow g(\mathbf{b}_N) \xrightarrow{P} g(\mathbf{b}). \tag{A.3}$$

Proof is given in Amemiya (1985, p. 79). Ruud (2000) presents a related result (see also Rao, 1973, p. 124) that lets the limit $\mathbf{b}$ be a random variable, at the expense of restricting $g(\cdot)$ to be continuous everywhere. Note that some authors instead refer to Theorem A.12 below as Slutsky's Theorem.

Theorem A.3 is one of the major reasons for the prevalence of asymptotic results versus finite-sample results in econometrics. It states a very convenient property that does not hold for expectations. For example, $\text{plim}(b_{1N}, b_{2N}) = (b_1, b_2)$ implies $\text{plim}(b_{1N}b_{2N}) = b_1 b_2$, whereas $\text{E}[b_{1N}b_{2N}]$ generally differs from $\text{E}[b_1]\text{E}[b_2]$.