Environmental Econometrics Using Stata

Christopher F. Baum
Department of Economics and School of Social Work
Boston College

Stan Hurn
School of Economics and Finance
Queensland University of Technology

A Stata Press Publication
StataCorp LLC
College Station, Texas

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

Print ISBN-10: 1-59718-355-5
Print ISBN-13: 978-1-59718-355-0
ePub ISBN-10: 1-59718-356-3
ePub ISBN-13: 978-1-59718-356-7
Mobi ISBN-10: 1-59718-357-1
Mobi ISBN-13: 978-1-59718-357-4

Library of Congress Control Number: 2021934557

Contents

Preface

There is no doubt that the environment is one of the greatest challenges faced by policymakers today. The key issues addressed by environmental sciences are often empirical. In many instances, very detailed, sizable datasets are available. Researchers in this field, including those in academe, research bodies, and government agencies, should have a solid understanding of the econometric tools best suited for analysis of these data.

Of course, there exist complex and expensive physical models of the environment that deal with many of the problems addressed in this book, such as pollution, temperature, greenhouse gas emissions, and sea levels to name but a few. However, it is becoming increasingly clear, through the increased involvement of econometricians in environmental issues that reduced-form models have a role to play not only in modeling environmental phenomena but also in producing point and density forecasts. In short, successful environmental modeling does not necessarily require a structural model, but it does require that the econometrics underlying the reduced-form approaches is competently done. This provides the essential raison d'être for the book.

This book is designed to introduce environmental researchers to a broad range of econometric techniques that can be effectively applied to environmental data. The study of environmental issues is inherently interdisciplinary, encompassing the physical sciences, economics, sociology, political science, and public health. Researchers in these fields are likely to have some statistical training, and an understanding of basic statistical concepts is presumed. The development of modern econometrics, coupled with increasing computational capability to process sizable datasets, has broadened our ability to study environmental data using powerful analytical and graphical tools.

Although our focus is on applied econometric techniques appropriate for the analysis of environmental data, we expect this book to be widely used. We believe that the potential audience includes economists at the undergraduate, graduate, and professional levels in academia, research institutes, consulting firms, government agencies, and international organizations. Our approach provides a gentle introduction to the most widely used econometric tools, which should serve to address the needs of those who may have only seen econometrics at an undergraduate level, such as those in public policy programs. We not only emphasize how to fit models in Stata but also highlight the need for using a wide range of diagnostic tests to validate the results of estimation and subsequent policy conclusions. This emphasis on careful, reproducible research should be appreciated by academic and non-academic researchers who are seeking to produce credible, defensible conclusions about key issues in environmental science.

Although appendix A provides a brief guide to using Stata effectively, thus book assumes that the reader is familiar with Stata's command line interface and elementary concepts of Stata programming such as do-files and data management facilities. An understanding of basic linear regression techniques will also be helpful but is not essential, because the book covers the basic building blocks of modern econometrics. More advanced econometric methods are also introduced, interspersing presentation of the underlying theory with clear examples of their employment on environmental data. In contrast with many existing econometric textbooks that deal mainly with the theoretical properties of estimators and test statistics, this book addresses the implementation issues that arise in the computational aspects of applied econometrics. The computer code that is provided will also help to bridge the gap between theory and practice so that the reader, as a result, can build on the code and tailor it to more challenging applications.

Organization

Although not specifically designated as such, the material presented in this book falls naturally into two parts. Chapters 1 to 8 provide a first course in applied environmental econometrics. These chapters cover the basic building blocks upon which the rest of the book is based, including the usual regression framework taught in standard econometric courses but always related to the modeling of environmental data. Chapter 2 describes the workhorse of applied econometrics, the linear regression model, while chapter 3 covers additional important estimation methods beyond the simple least-squares method. Chapter 4 extends the single-equation model to include dynamic components, while chapter 5 considers multiple time-series models, particularly vector autoregression and structural vector autoregression. The next two chapters develop the tools to deal with nonstationary data. Chapter 6 presents a range of tests for nonstationarity, known as unit-root tests. Chapter 7 discusses the extension of nonstationarity to deal with multiple time series and the idea of cointegrated systems. The last chapter in the first part of the book is chapter 8, which deals with forecasting methods and evaluation of forecast accuracy. Our philosophy is to make the treatment accessible by avoiding, wherever possible, the use of matrix algebra and potentially confusing notation. Where the use of this kind of notation is unavoidable, our intention is to provide as much intuition as possible.

The second part of the book comprises chapters 9 to 15 and relates to important econometric methods that may be of particular interest to empirical environmental studies. These chapters could form the core of a second course in applied environmental econometrics, and the level of difficulty steps up slightly. The first three chapters deal with techniques aimed at dealing with the nonlinear behavior that characterizes many environmental data series. Consequently, chapter 9 presents unobserved component models that decompose a given time series into its unobserved components, chapter 10 covers models that exhibit fundamental nonlinearity in mean, such as threshold models and Markov switching models, and chapter 11 deals with models that are nonlinear in variance and display what is known as volatility clustering. The remaining four

chapters are perhaps best described as topics in applied environmental econometrics. Chapter 12 illustrates selected models for longitudinal data, taking advantage of the multiple measurements of environmental series, such as climate conditions. Chapter 13 is concerned with models for data that are measured at different geographical locations and covers the estimation of models that are characterized by spatial effects. Chapter 14 presents a selection of limited dependent variable models, focusing on the modeling of willingness to pay for environmental preservation and mitigation. Chapter 15 presents models of fractional integration and cointegration, which were first studied in hydrology and biological processes.

The Stata code and datasets to reproduce all the examples in the book are available from a companion website. One of the features of this book is that each chapter has several nontrivial exercises that not only reinforce the material covered in the chapter but also extend it. Code to solve the exercises at the end of each chapter is also available.

6 Testing for nonstationarity

An important property of many environmental time series is that they exhibit strong trends. Series that are characterized by trending behavior are referred to as being nonstationary. The presence of nonstationarity in the time-series representation of a variable has important implications for both the econometric method used and the economic interpretation of the model in which that variable appears. This chapter focuses on identifying and testing for nonstationarity in environmental time series, while chapter 7 deals with modeling with nonstationary variables.

A variable y_t is said to be stationary if its distribution, or some important aspects of its distribution, is constant over time. There are two commonly used definitions of stationarity, known as weak (or covariance) and strong (or strict) stationarity. It is the former that will be of primary interest here. A process is weakly stationary if both the population mean and the population variance are constant over time and if the covariance between two observations is a function only of the distance between them and not of time.[1] Although the concept of nonstationarity is often mentioned in terms of a unit-root process, it is important to note that any time series with a time-varying mean or time-varying variance is a nonstationary process.

6.1 Per capita CO_2 emissions

The Earth absorbs energy from the sun and emits energy into space with the difference between incoming and outgoing radiation being known as radiative forcing. When incoming energy is greater than outgoing energy, positive radiative forcing will cause the planet to warm. Natural phenomena that contribute to radiative forcing include changes in the sun's energy output, changes in Earth's orbit, and volcanic activity. Anthropogenic (or human-caused) radiative forcing includes emissions of heat-trapping greenhouse gases.

It is well known that per capita CO_2 emissions, which are an important component of anthropogenic radiative forcing, have grown over time. Figure 6.1 plots annual per capita CO_2 emissions for the United States for the period 1870 to 2000. The data are provided by the United States Department of Energy through its Carbon Dioxide Information Analysis Center at Oak Ridge National Laboratory. An observation is an annual number giving national emissions in metric tons of carbon from fossil fuel burning, cement manufacturing, and gas flaring.

1. Strict stationarity is a stronger requirement than that of weak stationarity and pertains to all the moments of the distribution, not just the first two.

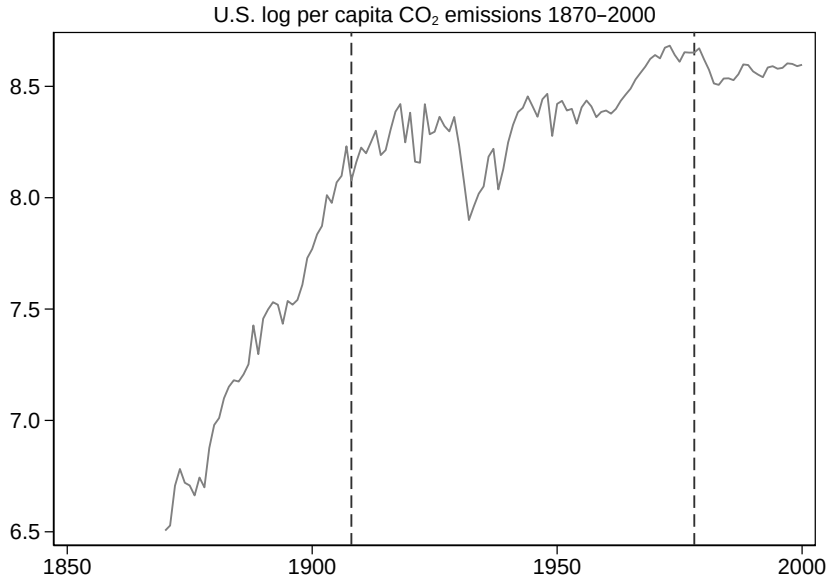

Figure 6.1. Annual per capita CO_2 emissions for the United States for the period 1870 to 2000, with three phases of emission development demarcated by the vertical lines

Per capita CO_2 emissions in industrialized countries are often characterized in terms of three phases as in Lanne and Liski (2004). The first phase (1870–circa 1908) was that of fast growth of per capita emissions because early industrialization and development in general was primarily dependent on coal. The second phase (circa 1908–circa 1978) was characterized by slower growth due to the shift from solid to nonsolid fuels (from coal to oil and gas). This diversification in national fuel compositions was partly induced by local but CO_2-related pollution problems and technological progress associated with the demand for higher energy density fuels. The third phase followed the oil price shocks of the 1970s, which may have permanently changed the structure of emissions from fossil fuels and possibly led to downward sloping per capita emission trends.

One of the important questions of environmental econometrics is to assess if this growth is consistent with deterministic behavior with respect to time or if the dependence on time is stochastic. A major challenge in time-series modeling is determining the source of nonstationary behavior, which may be due to either a deterministic trend or a stochastic trend. A deterministic trend is a nonrandom function of time

$$y_t = \alpha + \delta t + v_t \tag{6.1}$$

in which t in this instance is a simple time trend taking integer values from 1 to T and v_t is an independently and identically distributed (i.i.d.) random variable. In this model, shocks to the system have a transitory effect in that the process always reverts to its mean of $\alpha + \delta t$. This implies that removing the deterministic trend from y_t will produce a stationary series. A series that is stationary once its deterministic trend has been removed is therefore called a *trend-stationary process*.

By contrast, a time series that exhibits a stochastic trend is random and varies over time, for example,

$$y_t = \alpha + y_{t-1} + v_t \tag{6.2}$$

which is known as a random walk with drift model. In this model, the best guess for the next value of series is the current value plus some constant rather than a deterministic mean value. This is a generalization of the pure random-walk process, in which α would not appear. As a result, this kind of model is also called a "local trend" or "local level" model. The appropriate course of action here is to difference the data to obtain a stationary series as follows:

$$\Delta y_t = \alpha + v_t$$

A process that has a stochastic trend that is removed by differencing the series is known as a *difference-stationary process*.[2]

If the behavior of per capita CO_2 is deterministic with respect to time, then the emissions series may be appropriately detrended. If this is not the case, then a new set of econometric tools dealing with fundamentally nonstationary variables must be used. However, distinguishing trend stationarity from difference stationarity is a difficult task that is made more so by the practical modeling problems encountered in the data. There are at least two problems evident in CO_2 emissions data plotted in figure 6.1.

1. It is obvious from the plot that a simple linear deterministic trend is unlikely to be an adequate representation of any deterministic behavior with respect to time.

2. Related to this point is the idea that the series may be characterized by structural breaks consistent with the phases of development of an industrialized country.

2. Of course, first differencing, a series that has a linear deterministic trend, such as (6.1) will automatically remove this trend. Why, then, is the solution not simply to difference all series, irrespective of whether they are trend or difference stationary? In first difference form, (6.1) becomes

$$\Delta y_t = \delta + v_t - v_{t-1}$$

so the process of taking the first difference introduces a moving average error term that has a coefficient of 1 on the lagged error. The nonstationarity has been transferred from the linear trend to a more complex form of nonstationarity in the error structure. This is known as overdifferencing, which introduces several econometric problems and is best avoided.

6.2 Unit roots

Consider again the random walk with drift model in (6.2). Recursively substituting for the lagged value of y_t on the right-hand side yields

$$\begin{aligned} y_t &= 2\alpha + y_{t-2} + v_t + v_{t-1} \\ y_t &= 3\alpha + y_{t-3} + v_t + v_{t-1} + v_{t-2} \\ \vdots &= \quad \vdots \quad\quad \vdots \\ y_t &= t\alpha + y_0 + v_t + v_{t-1} + v_{t-2} + \cdots + v_1 \end{aligned}$$

This demonstrates that y_t is the summation of all past disturbances. The element of summation of the disturbances in a nonstationary process yields an important concept—the order of integration of a series. A process is said to be integrated of order d, denoted by $I(d)$, if it can be rendered stationary by differencing d times. That is, y_t is nonstationary but $\Delta^d y_t$ is stationary. Accordingly, a process is said to be integrated of order one, denoted by $I(1)$, if it can be rendered stationary by differencing once; y_t is nonstationary but $\Delta y_t = y_t - y_{t-1}$ is stationary. If $d = 2$, then y_t is $I(2)$ and must be differenced twice to achieve stationarity.

A series that is $I(1)$ is also said to have a unit root, and tests for nonstationarity are commonly known as tests for unit roots. Consider the general nth order autoregressive (AR) process

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_n y_{t-n} + v_t$$

This may be rewritten using the lag operator, L, defined in chapter 4 so that

$$y_t = \phi_1 L y_t + \phi_2 L^2 y_t + \cdots + \phi_n L^n y_t + v_t$$

or

$$\Phi(L)\, y_t = v_t$$

where

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_n L^n$$

is a polynomial in the lag operator of order n. The roots of this polynomial are the values of L, which satisfy the equation

$$1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_n L^n = 0$$

If all the roots of this equation are greater than 1 in absolute value, then y_t is stationary. If, on the other hand, any of the roots lie within the unit circle, then y_t is nonstationary.

For the AR(1) model

$$(1 - \phi_1 L)\, y_t = v_t$$

the roots of the equation

$$1 - \phi_1 L = 0$$

are of interest. The single root of this equation is given by $L^* = 1/\phi_1$, and the root is greater than 1 if and only if $|\phi_1| < 1$. If this is the case, then the AR(1) process is stationary. If, on the other hand, the root of the equation is unity, then $|\phi_1| \geq 1$ and the AR(1) process is nonstationary.

In the AR(2) model

$$\left(1 - \phi_1 L - \phi_2 L^2\right) y_t = v_t$$

there are two unit roots, corresponding to the roots of the equation

$$1 - \phi_1 L - \phi_2 L^2 = 0$$

and y_t will have a unit root if either of the roots is unity. In a AR(2) model, the solution could also consist of a complex conjugate pair of roots. In that case, the modulus of the complex pair must lie inside the unit circle if the series is to be stationary. The presence of complex roots introduces cyclical behavior in the series.

To illustrate these alternatives, the following series are simulated for three AR(2) processes:

1. $y_t = 0.9y_{t-1} - 0.2y_{t-2} + v_t$
2. $y_t = 0.8y_{t-1} - 0.5y_{t-2} + v_t$
3. $y_t = 1.8y_{t-1} - 0.8y_{t-2} + v_t$

The first process has two real roots outside the unit circle, implying that both (φ_1, φ_2) are less than 1 in absolute value. The series is stationary as shown in figure 6.2.

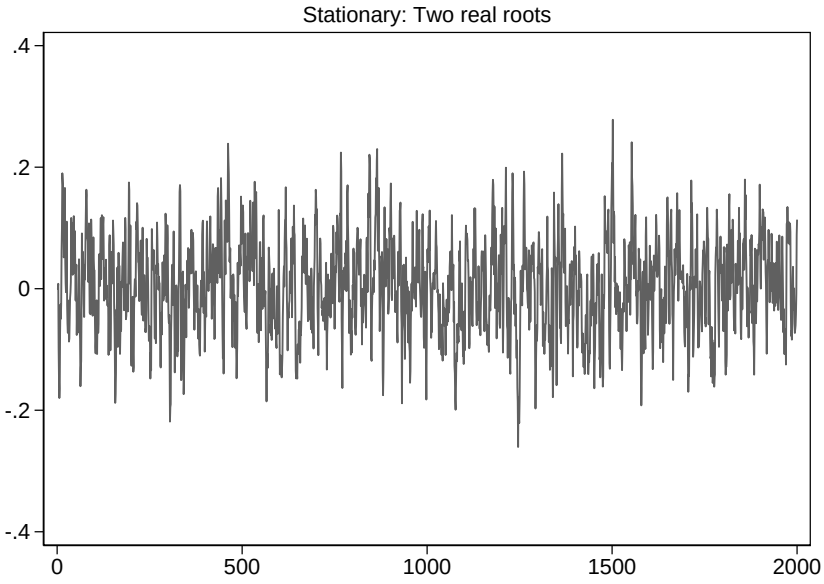

Figure 6.2. AR(2) model with two real roots

The second process has a pair of complex roots, with its modulus inside the unit circle, and also exhibits stationary behavior, as shown in figure 6.3.

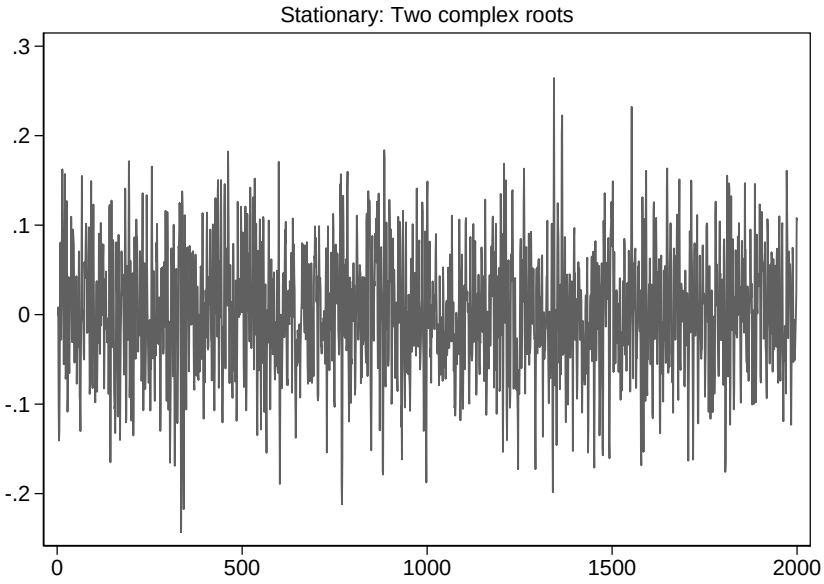

Figure 6.3. AR(2) model with complex roots

The third process has one unit root because the sum of its lag coefficients is 1. Its second root of 1.44 is outside the unit circle. The presence of a unit root defines an $I(1)$ process, drifting arbitrarily far from its starting values, as shown in figure 6.4.

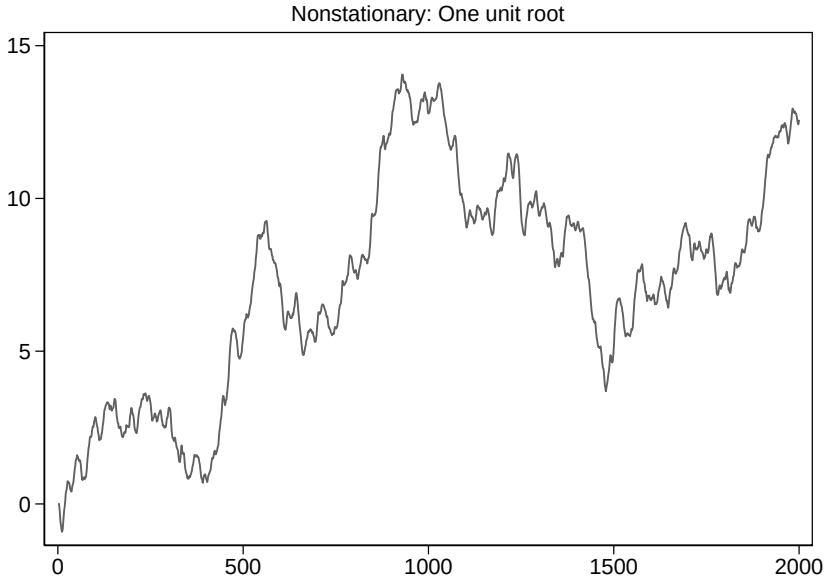

Figure 6.4. AR(2) model with one unit root

In the event of $\phi_1 = 2$ and $\phi_2 = -1$, then both absolute roots of the equation are one. The y_t series has two unit roots and is therefore $I(2)$. We do not illustrate a process with two unit roots such as that described above with coefficients of $(2, -1)$, because its behavior is monotonic, increasing without bound.

14 Discrete dependent variables

In most of the models previously discussed in this book, the dependent variable is assumed to be a continuous random variable. There are several situations where this assumption is inappropriate, and alternative classes of models must be specified. This often calls for a model of discrete choice in which the response variable is restricted to a Boolean or binary choice, indicating that a particular course of action was or was not selected. In other models, the dependent variable may take only integer values such as the ordered values on a Likert scale.[1] Alternatively, the dependent variable may appear to be a continuous variable with several responses at a threshold value. For instance, the response to the question of how much you are willing to pay for environmental conservation will be bounded by zero, making these kinds of dependent variables unsuitable for modeling by means of linear regression methods.

There are two important types of discrete dependent variables that will not be covered in this chapter. These are categorical data, in which the dependent variable falls into one of several mutually exclusive but unordered categories, and count data, where the dependent variable takes only nonnegative values. In both of these cases, just as with the other limited dependent variable models, linear regression is not an appropriate estimation technique. Interested readers are referred to Cameron and Trivedi (2010, 2013) for excellent treatments of discrete dependent variable models.

The discrete time-series models that will be discussed here are logit, probit, ordered probit, and Tobit regression models. This choice should not be interpreted as suggesting that these are the only important models for discrete dependent variables but rather that the ideas and methods in these models are broadly compatible and lend themselves to being modeled in one chapter. From a purely practical perspective, obtaining one dataset that facilitates the discussion of a broader set of discrete dependent variable models is difficult.

14.1 Humpback whales

Each year between April and November, eastern humpback whales migrate north for about 10,000 kilometers from their feeding grounds in Antarctic waters to subtropical waters where they mate and give birth. Australia's eastern coastline comes alive with the spectacular acrobatic displays of humpback whales. Before commercial whaling, an

1. A Likert scale is a rating scale used in surveys that is aimed at ascertaining how people feel about something. Typically, responders are asked to indicate their agreement or disagreement with a statement typically in five or seven points.

estimated population of around 40,000 humpback whales migrated along the east coast of Australia.

Shortly after European colonization, whaling and the export of whale products became Australia's first primary industry. Australian and New Zealand whalers of the early 19th century hunted from small boats, towing their catch back for processing at shore stations. The development of harpoon guns, explosive harpoons, and steam-driven whaling ships later that century made large-scale commercial whaling so efficient that many whale species were overexploited in the 20th century and came very close to extinction. It is believed that up to 95% of the east coast population of humpbacks was killed in the decade from 1952 to 1962. By 1963, when whaling ended, there may have only been 500 whales left. This disastrous period in the history of the eastern humpback whale is graphically illustrated in figure 14.1, where the annual catch for Australian and New Zealand is plotted, representing a grand total of 14399 eastern humpbacks; see Jackson et al. (2008).

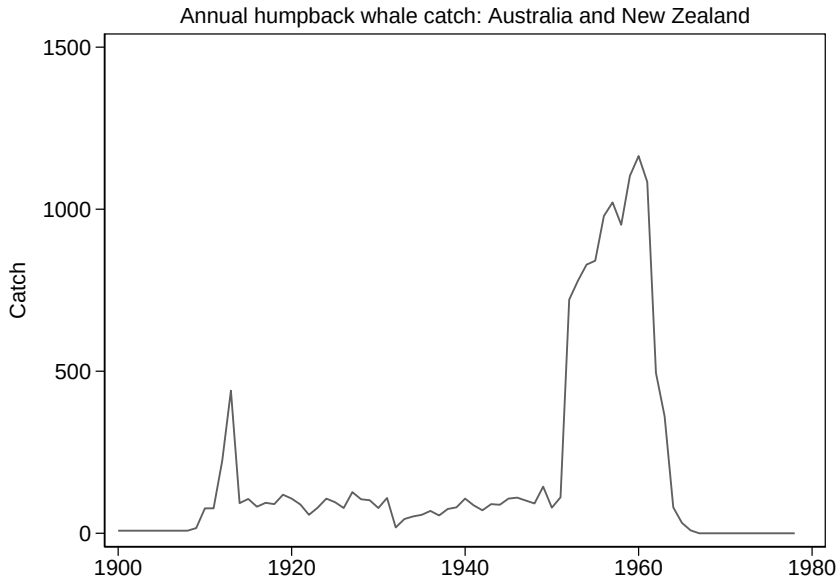

Figure 14.1. Time-series plot of the estimated annual catch of humpback whales for Australia and New Zealand

The International Whaling Commission banned humpback whaling in the Southern Hemisphere in 1963, and a series of research and monitoring programs has allowed accurate estimates to be made of the growth rate and size of the humpback whale population that migrates along the east coast of Australia. Left alone, the remnants of the population staged a miraculous recovery, increasing in number by around 10% each year. In 2013, the total population estimate was around 19,000 and is currently estimated to be back to about 90% of the prewhaling population.

The recovery of the humpback population has contributed significantly to the rapid growth of Australia's whale-watching industry. During their annual migration, humpbacks attract thousands of visitors to coastal towns along the subtropical east coast of Australia.

14.2 The data

The dataset used in this chapter is the result of a survey conducted in the Queensland town of Hervey Bay by a group of academics seeking to measure the willingness to pay for whale conservation efforts.[2] The survey was conducted in 2000 and had 701 respondents who were surveyed right after returning from a whale-watching trip in Hervey Bay.

The discrete dependent variables for the econometric methods demonstrated in this chapter are constructed from the answers to the following three questions in the survey.

1. Would you be willing to have your take home income reduced by $2 per week for the next 10 years to protect and conserve whales that come to breed in Australian waters?

   Yes ☐ No ☐

2. Following your visit to Hervey Bay, are you willing to pay

   More ☐ Less ☐ Same ☐

   for humpback whale protection and conservation as before your visit?

3. To protect and conserve humpback whales that come to Australia to breed, what is the **maximum** amount you would be willing to pay per week for the next 10 years?

   Aus $ ..... per week

2. The data used in this chapter were kindly made available by Professor Clevo Wilson of the Queensland University of Technology.

The answers to these questions give rise to different kinds of dependent variables and fundamental differences in the kinds of econometric models that are appropriate in each case. The answer to the first question creates a binary dependent variable ($\texttt{PayConserve}_k$), the answer to the second defines an ordered dependent variable ($\texttt{wtp}_k$), and the answer to the third should be considered as a censored dependent variable ($\texttt{Max\_wtp}_k$). Each of these is now examined in more detail.

Using the survey data, the results from the first 20 respondents are as follows:

```
. use http://www.stata-press.com/data/eeus/whalesdata
. list PayConserve wtp Max_wtp in 1/20
```

	PayCon~e	wtp	Max_wtp
1.	1	Less	0
2.	0	More	0
3.	0	More	0
4.	1	More	0
5.	0	More	0
6.	1	Same	1
7.	1	Less	2
8.	0	More	0
9.	1	Same	5
10.	0	More	0
11.	0	More	0
12.	1	Same	5
13.	1	Same	2
14.	0	More	0
15.	1	Same	2
16.	1	Same	4
17.	0	More	0
18.	1	Same	10
19.	0	More	0
20.	1	More	0

The binary dependent variable, $\texttt{PayConserve}_k$, is created by applying the rule

$$\texttt{PayConserve}_k = \begin{cases} 1 & \text{yes} \\ 0 & \text{no} \end{cases}$$

The linear probability model and the binomial logit or probit models are appropriate for this kind of dependent variable.

The second column of data is an example of an ordered dependent variable. Here the question about willingness to pay is phrased slightly differently. After the whale-watching trip, the respondents were asked to say whether they would now be willing to pay less, the same, or more for whale conservation. The ordered data are created by applying the rule

$$\texttt{wtp}_k = \begin{cases} 1 & \text{willing to pay less than before} \\ 2 & \text{willing to pay the same as before} \\ 3 & \text{willing to pay more than before} \end{cases}$$

These data are obviously more informative than the simple binary variable, and the appropriate econometric models are the ordered probit or logit models.

In the case of the censored model, respondents were asked to indicate how much per week over a 10-year period they were willing to contribute to whale conservation. Let $\widetilde{y}_t$ now represent the amount offered, a variable that is censored at zero, so the censored dependent variable y_t is created as follows:

$$\texttt{Max\_wtp}_k = \begin{cases} \widetilde{y}_t & \widetilde{y}_t > 0 \\ 0 & \widetilde{y}_t \leq 0 \end{cases}$$

with the results recorded in the column headed censored. This model is a mixture of the full information model where $\widetilde{y}_t > 0$ and the binary model where $\widetilde{y}_t < 0$. This is the censored regression or Tobit model.

In terms of explanatory variables used in the study, the `describe` command produces the following output:

```
. describe
Contains data from http://www.stata-press.com/data/eeus/whalesdata.dta
  obs:           701
 vars:            15                          23 Oct 2018 14:30
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
Country         str15   %15s
SeeWhales       byte    %8.0g                 Have you seen humpback whales at
                                                Hervey Bay before
Age             byte    %8.0g                 Q3.1A Age of respondent
Gender          byte    %8.0g      GENDER     Q3.1G Gender of respondent
Education       byte    %8.0g      EDUCATIO   Q3.2 Highest qualification of
                                                respondent
Income_AUD      byte    %8.0g      INCOME_A   Q3.4.1 Household income in
                                                Australian dollars
Income          float   %9.0g                 Mid point
AdultWhales     byte    %8.0g                 Q4.1How many adult humpback whales
                                                did you see
YoungWhales     byte    %8.0g      YOUNGWHA   Q4.2 Did you see young humpback
                                                whales
Max_wtp         double  %10.0g                Q7.4 Maximum WTP/ week to conserve
                                                whales
wtp             byte    %8.0g      WTP        Q7.5 Following WW trip are you
                                                willing to pay for protection
PayConserve     byte    %9.0g                 Willingness to pay to conserve
                                                whales (1=Yes; 0=No)
foreign         byte    %9.0g
highered        byte    %9.0g
scaledWtp       float   %9.0g
-------------------------------------------------------------------------------
Sorted by:
```

The distribution of income, measured in Australian dollars, of the respondents reveals the following:

```
. tabulate Income_AUD
 Q3.4.1 Household
        income in
       Australian
          dollars |      Freq.     Percent        Cum.
------------------+-----------------------------------
Less than $20,000 |        138       20.47       20.47
$20,001 - $30,000 |         87       12.91       33.38
$30,001 - $40,000 |        116       17.21       50.59
$40,001 - $50,000 |         62        9.20       59.79
$50,001 - $60,000 |         80       11.87       71.66
$60,001 - $70,000 |         51        7.57       79.23
More than $70,000 |        140       20.77      100.00
------------------+-----------------------------------
            Total |        674      100.00
```

The distribution of income reflects the age distribution in figure 14.2. The large number of respondents earning less than $40,000 follows directly from the spike in the age distribution in the early to mid 20s. The large number of respondents above $70,000 is perhaps indicative that the survey prematurely "top-coded" the income scale. This is partly due to the fact that there were many visitors from the United States and the United Kingdom. In September 2000, the height of the whale-watching season, the Australian dollar was weak against these currencies, resulting in a bracket creep for these respondents.

About 70% of the respondents were Australian, and 60% had completed secondary education and had some sort of tertiary education. The age distribution of the respondents is shown in figure 14.2. The distribution is fairly typical of the age of visitors to Queensland with a relatively higher proportion than would be expected in the early to mid-20s (backpackers) and in the late 50s and early 60s (the gray nomads).

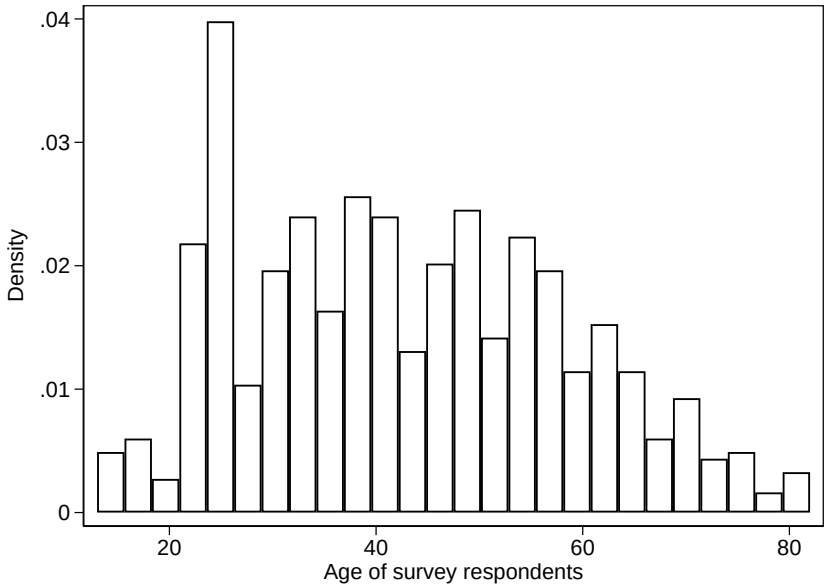

Figure 14.2. Histogram of age of the 701 respondents to the Hervey Bay whale conservation questionnaire

Finally, note that the `foreign` variable takes the value 1 if the respondent's nationality is not Australian and 0 otherwise. Also note the `highered` variable is a simplified version of `Education` that takes the value 1 if the respondent has education above the secondary level.