*Wooldridge, Introductory Econometrics*

## Chapter 10: Basic regression analysis with time series data

We now turn to the analysis of time series data. One of the key assumptions underlying our analysis of cross-sectional data will prove to be untenable when we consider time series data; thus, we separate out the issues of time series modelling from that of cross sections. How does time series data differ? First of all, it has a natural ordering, that of calendar time at some periodic frequency. Note that we are not considering here a dataset in which some of the variables are dated at a different point in time: e.g. a survey measuring this year's income, and (as a separate variable) last year's income. In time series data sets, the observations are dated, and thus we need to respect

their order, particularly if the model we consider has a **dynamic** specification (involving variables from more than one point in time). What is a time series? Merely a sequence of observations on some phenomenon observed at regular intervals. Those intervals may correspond to the passage of calendar time (e.g. annual, quarterly, monthly data) or they may reflect an economic process that is irregular in calendar time (such as business-daily data). In either case, our observations may not be available for every point in time (for instance, there are days when a given stock does not trade on the exchange).

A second important difference between cross-sectional and time series data: with the former, we can reaonably assume that the sample is drawn randomly from the appropriate population, and could conceive of one or many alternate samples constructed from the same population. In the case of time series data, we consider the sequence of events we have recorded

as a **realization** of the underlying process. We only have one realization available, in the sense that history played out a specific sequence of events. In an alternate universe, Notre Dame might have lost to BC this year, or the Red Sox might not have triumphed in the World Series. Randomness plays a role, of course, just as it does in cross-sectional data; we do not know what will transpire until it happens, so that time series data *ex ante* are random variables. We often speak of a time series as a **stochastic process**, or time series process, focusing on the concept that there is some mechanism generating that process, with a random component.

## Types of time series regression models

Models used in a time series context can often be grouped into those sharing common features. By far the simplest is a **static** model, such as

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + u_t \qquad (1)$$

We may note that this model is the equivalent of the cross-sectional regression model, with the $i$ subscript in the cross section replaced by $t$ in the time series context. Each observation is modeled as depending only on **contemporaneous** values of the explanatory variables. This structure implies that all of the interactions among the variables of the model are assumed to take place *immediately*: or, taking the frequency into account, within the same time period. Thus, such a model might be reasonable when applied to annual data, where the length of the observation interval is long enough to allow behavioral adjustments to take place. If we applied the same model to higher-frequency data, we might consider that assumption inappropriate; we might consider, for instance, that a tax cut would not be fully reflected by higher retail sales in the same month that it took effect. An example of such

a structure that appears in many textbooks is the static Phillips curve:

$$\pi_t = \beta_0 + \beta_1 UR_t + u_t \tag{2}$$

where $\pi_t$ is this year's inflation rate, and $UR_t$ is this year's unemployment rate. Stating the model in this form not only implies that the level of unemployment is expected to affect the rate of inflation (presumably with a negative sign), but also that the entire effect of changes in unemployment will be reflected in inflation within the observation interval (e.g. one year).

In many contexts, we find a static model inadequate to reflect what we consider to be the relationship between explanatory variables and those variables we wish to explain. For instance, economic theory surely predicts that changes in interest rates (generated by monetary policy) will have an effect on firms' capital investment spending. At lower interest rates,

firms will find more investment projects with a positive expected net present value. But since it takes some time to carry out these projects–equipment must be ordered, delivered, and installed, or new factories must be built and equipped–we would not expect that quarterly investment spending would reflect the same quarter's (or even the previous quarter's) interest rates. Presumably interest rates affect capital investment spending with a lag, and we must take account of that phenomenon. If we were to model capital investment with a static model, we would be omitting relevant explanatory variables: the prior values of the causal factors. These omissions would cause our estimates of the static model to be biased and inconsistent. Thus, we must use some form of **distributed lag** model to express the relationship between current and past values of the explanatory variables and the outcome. Distributed lag models may take a finite number

of lagged values into account (thus the Finite Distributed Lag model, or FDL) or they may use an infinite distributed lag: e.g. all past values of the $x$ variables. When an infinite DL model is specified, some algebraic sleight-of-hand must be used to create a finite set of regressors.

A simple FDL model would be

$$f_t = \beta_0 + \beta_1 pe_t + \beta_2 pe_{t-1} + \beta_3 pe_{t-2} + u_t \quad (3)$$

in which we consider the fertility rate in the population as a function of the personal exemption, or child allowance, over this year and the past two years. We would expect that the effect of a greater personal exemption is positive, but realistically we would not expect the effect to be (only) contemporaneous. Given that there is at least a 9-month lag between the decision and the recorded birth, we would expect such an effect (if it exists) to be largely

concentrated in the $\beta_2$ and $\beta_3$ coefficients. Indeed, we might consider whether additional lags are warranted. In this model, $\beta_1$ is the **impact effect**, or **impact multiplier** of the personal exemption, measuring the contemporaneous change. How do we calculate $\partial f/\partial pe$? That (total) derivative must be considered as the effect of a one-time change in $pe$ that raises the exemption by one unit and leaves it permanently higher. It may be computed by evaluating the **steady state** of the model: that with all time subscripts dropped. Then it may be seen that the total effect, or
**long-run multiplier**, of a permanent change in $pe$ is $(\beta_1 + \beta_2 + \beta_3)$. In this specification, we presume that there is an impact effect (allowing for a nonzero value of $\beta_1$) but we are imposing the restriction that the entire effect will be felt within the two year lag. This is testable, of course, by allowing for additional lag terms in the model, and testing for their joint significance. However the analysis of individual

lag coefficients is often hampered–especially at higher frequencies such as quarterly and monthly data–by high autocorrelation in the series. That is, the values of the series are closely related to each other over time. If this is the case, then many of the individual coefficients in a FDL regression model may not be distinguishable from zero. This does not imply, though, that the sum of those coefficients (i.e. the long run multiplier) will be imprecisely estimated. We may get a very precise value for that effect, even if its components are highly intercorrelated.

One additional concern that will apply in estimating FDL models, especially when the number of observations is limited. Each lagged value included in a model results in the loss of one observation in the estimation sample. Likewise, the use of a first difference ($\Delta y_t \equiv y_t - y_{t-1}$) on either the left or right

side of a model results in the loss of one observation. If we have a long time series, we may not be too concerned about this; but if we were working with monthly data, and felt it appropriate to consider 12 lags of the explanatory variables, we would lose the first year of data to provide these starting values. Computer programs such as Stata may be set up to recognize the time series nature of the data (in Stata, we use the `tsset` command to identify the date variable, which must contain the calendar dates over which the data are measured), and construct lags and first differences taking these constraints into account (for instance, a lagged value of a variable will be set to a missing value where it is not available). In Stata, once a dataset has been established as time series, we may use the operators `L.`,`D.` and `F.` to refer to the lag, difference or lead of a variable, respectively: so `L.gdp` is last period's `gdp`, `D.gdp` is the first difference, and `F.gdp` is

next year's value. These operators can also consider higher lags, so `L2.gdp` is the second lag, and `L(1/4).gdp` refers to the first four lags, using standard Stata "numlist" notation (`help numlist` for details).

## Finite sample properties of OLS

How must we modify the assumptions underlying OLS to deal with time series data? First of all, we assume that there is a linear model linking $y$ with a set of explanatory variables, $\{x_1 ... x_k\}$, with an additive error $u$,for a sample of $t = 1, ..., T$. It is useful to consider the explanatory variables as being arrayed in a matrix:

$$X = \begin{matrix} x_{1,1} & \cdots & x_{1,k} \\ x_{2,1} & \cdots & x_{2,k} \\ \vdots & \cdots & \vdots \\ x_{T,1} & \cdots & x_{n,k} \end{matrix}$$

where, like a spreadsheet, the rows are the observations (indexed by time) and the columns are the variables (which may actually be dated differently: e.g. $x_2$ may actually be the lag of $x_1$, etc.) To proceed with the development of the finite sample properties of OLS, we assume:

**Proposition 1** *For each* $t$, $E(u_t|X) = 0$, *where* $X$ *is the matrix of explanatory variables.*

This is a key assumption, and quite a strong one: it states not only that the error is contemporaneously uncorrelated with each of the explanatory variables, but also that the error is assumed to be uncorrelated with elements of $X$ at every point in time. The weaker statement of **contemporaneous exogeneity**, $E(u_t|x_{t,1}, x_{t,2}, ..., x_{t,k}) = 0$ is analogous to the assumption that we made in the cross-sectional context. But this is a stronger assumption, for

it states that the elements of $X$, past, present, and future, are independent of the errors: or that the explanatory variables in $X$ are **strictly exogenous.** It is important to note that this assumption, by itself, says nothing about the correlations over time among the explanatory variables (or their correlations with each other), nor about the possibility that successive elements of $u$ may be correlated (in which case we would say that $u$ is **autocorrelated**). The assumption only states that the distributions of $u$ and $X$ are independent.

What might cause this assumption to fail? Clearly, omitted variables and/or measurement error are likely causes of a correlation between the regressors and errors. But in a time series context there are other likely suspects. If we estimate a static model, for instance, but the true relationship is **dynamic**: in which lagged values of some of the explanatory variables also

have direct effects on $y$—then we will have a correlation between contemporaneous $x$ and the error term, since it will contain the effects of lagged $x$, which is likely to be correlated with current $x$. So this assumption of strict exogeneity has strong implications for the correct specification of the model (in this case, we would need to specify a FDL model). It also implies that there cannot be correlation between current values of the error process and *future* $x$ values: something that would be likely in a case where some of the $x$ variables are policy instruments. For instance, consider a model of farmers' income, dependent on (among other factors) on government price supports for their crop. If unprecedented shocks (such as a series of droughts), which are unpredictable and random effects of weather on farmers' income, trigger an expansion of the government price support program, then the errors today are correlated with future $x$ values.

The last assumption we need is the standard assumption that the columns of $X$ are *linearly independent*: that is, there are no exact linear relations, or **perfect collinearity**, among the regressors.

With these assumptions in hand, we can demonstrate that the OLS estimators are unbiased, both conditional on $X$ and unconditionally. The random assumption that allowed us to prove unbiasedness in the cross-sectional context has been replaced by the assumption of strict exogeneity in the time series context. We now turn to the interval estimates. As previously, we assume that the error variance, conditioned on $X$, is homoskedastic:
$Var(u_t|X) = Var(u_t) = \sigma^2, \forall t.$
In a time series context, this assumption states that the error variance is constant over time, and in particular not influenced by the $X$ variables. In some cases, this may be quite unrealistic. We now add an additional assumption, particular to time series analysis: that

there is no **serial correlation** in the errors: $Cov(u_t, u_s|X) = Cov(u_t, u_s) = 0, \forall t \neq s$.
This assumption states that the errors are not **autocorrelated**, or correlated with one another, so that there is no systematic pattern in the errors over time. This may clearly be violated, if the error in one period (for instance, the degree to which the actual level of $y$ falls short of the desired level) is positively (or negatively) related to the error in the previous period. Positive autocorrelation can readily arise in a situation where there is partial adjustment to a discrepancy, whereas negative autocorrelation is much more likely to reflect "overshooting," in which a positive error (for instance, an overly optimistic forecast) is followed by a negative error (a pessimistic forecast). This assumption has nothing to do with the potential autocorrelation within the $X$ matrix; it only applies to the error process. Why is this assumption only relevant for time

series? In cross sections, we assume random sampling, whereby each observation is independent of every other. In time series, the sequence of the observations makes it likely that if independence is violated, it will show up in successive observations' errors.

With these additional assumptions, we may state the Gauss-Markov theorem for OLS estimators of a time series model (OLS estimators are BLUE), implying that the variances of the OLS estimators are given by:

$$Var(b_j|X) = \frac{\sigma^2}{\left[SST_j\left(1 - R_j^2\right)\right]} \tag{4}$$

where $SST_j$ is the total sum of squares of the $j^{th}$ explanatory variable, and $R_j^2$ is the $R^2$ from a regression of variable $x_j$ on the other elements of $X$. Likewise, the unknown parameter $\sigma^2$ may be replaced by its consistent estimate,

$s^2 = \frac{SSR}{n-k-1}$, identical to that discussed previously.

As in our prior derivation, we will assume that the errors are normally distributed: $u \sim N(0, \sigma^2)$. If the above assumptions hold, then the standard $t-$statistics and $F-$statistics we have applied in a cross-sectional context will also be applicable in time series regression models.

**Functional form, dummy variables, and index numbers**

We find that a logarithmic transformation is very commonly used in time series models, particularly with series that reflect stocks, flows, or prices (rather than rates). Many models are specified with the first difference of $\log(y)$, implying that the dependent variable is the growth rate of $y$. Dummy variables are also very useful to test for **structural change**. We

may have *a priori* information that indicates that unusual events were experienced in particular time periods: wars, strikes, or presidential elections, or a market crash. In the context of a dynamic model, we do not want to merely exclude those observations, since that would create episodes of missing data. Instead, we can "dummy" the period of the event, which then allows for an intercept shift (or, with interactions, for a slope shift) during the unusual period. The tests for significance of the dummy coefficients permit us to identify the importance of the period, and justify its special treatment. We may want to test that the relationship between inflation and unemployment (the "Phillips curve") is the same in Republican and Democratic presidential administrations; this may readily be done with a dummy for one party, added to the equation and interacted to allow for a slope change between the two parties' equations. Dummy variables are

also used widely in financial research, to conduct **event studies**: models in which a particular event, such as the announcement of a takeover bid, is hypothesized to trigger "abnormal" returns to the stock. In this context, high-frequency (e.g. daily) data on stock returns are analyzed, with a dummy set equal to 1 on and after the date of the takeover bid announcement. A test for the significance of the dummy coefficient allows us to analyze the importance of this event. (These models are explicitly discussed in EC327, Financial Econometrics).

Creation of these dummies in Stata is made easier by the `tin()` function (read: tee-in). If the data set has been established as a time series via `tsset`, you may refer to natural time periods in generating new variables or specifying the estimation sample. For instance,

`gen prefloat = (tin(1959q1,1971q3))` will generate a dummy for that pre-Smithsonian period, and a model may be estimated over a subset of the observations via `regress ... if tin(1970m1,1987m9)`.

In working with time series data, we are often concerned with series measured as index numbers, such as the Consumer Price Index, GDP Deflator, Index of Industrial Production, etc. The price series are often needed to generate real values from nominal magnitudes. The usual concerns must be applied in working with these index number series, some of which have been rebased (e.g. from 1982=100 to 1987=100) and must be adjusted accordingly for a new base period and value. Interesting implications arise when we work with "real" magnitudes, expressed in logs: for instance, labor supply is usually modelled as depending on the real wage, $\left(\frac{w}{p}\right)$. If we express these variables in logs,

the log of the real wage becomes $\log w - \log p$. Regressing the log of hours worked on a single variable, $(\log w - \log p)$, is a restricted version of a regression in which the two variables are entered separately. In that regression, the coefficients will almost surely differ in their absolute value. But economic theory states that only the real wage should influence workers' decisions; they should not react to changes in its components (e.g. they should not be willing to supply more hours of labor if offered a higher nominal wage that only makes up for a decrease in their purchasing power).

## Trends and seasonality

Many economic time series are **trending**: growing over time. One of the reasons for very high $R^2$ values in many time series regressions is the common effect of time on many of the variables considered. This brings a challenge to

the analysis of time series data, since when we estimate a model in which we consider the effect of several causal factors, we must be careful to account for the co-movements that may merely reflect trending behavior. Many macro series reflect upward trends; some, such as the cost of RAM for personal computers, exhibit strong downward trends. We can readily model a **linear trend** by merely running a regression of the series on $t$, in which the slope coefficient is then $\partial y/\partial t$. To create a time trend in Stata, you can just `generate t = _n`, where `_n` is the observation number. It does not matter where a trend starts, or the units in which it is expressed; a trend is merely a series that changes by a fixed amount per time period. A linear trend may prove to be inadequate for many economic series, which we might expect on a theoretical basis to exhibit constant growth, not constant increments. In this case, an **exponential trend** may readily be estimated (for

strictly positive $y$) by regressing $\log y$ on $t$. The slope coefficient is then a direct estimate of the percentage growth rate per period. We could also use a polynomial model, such as a **quadratic time trend**, regressing the level of $y$ on $t$ and $t^2$.

Nothing about trending economic variables violates our basic assumptions for the estimation of OLS regression models with time series data. However, it is important to consider whether significant trends exist in the series; if we ignore a common trend, we may be estimating a **spurious regression**, in which both $y$ and the $X$ variables appear to be correlated because of the influence on both of an omitted factor, the passage of time. We can readily guard against this by including a time trend (linear or quadratic) in the regression; if it is needed, it will appear to be a significant determinant of $y$. In some cases, inclusion of a

time trend can actually highlight a meaningful relationship between $y$ and one or more $x$ variables: since their coefficients are now estimates of their co-movement with $y$, *ceteris paribus*: that is, net of the trend in $y$.

We may link the concept of a regression inclusive of trend to the common practice of analyzing **detrended** data. Rather than regressing $y$ on $X$ and $t$, we could remove the trend from $y$ and each of the variables in $X$. How? Regress each variable on $t$, and save the residuals (if desired, adding back the original mean of the series). This is then the detrended $y$, call it $y^*$, and the detrended explanatory variables $X^*$ (not including a trend term). If we now estimate the regression of $y^*$ on $X^*$, we will find that the slope coefficients' point and interval estimates are exactly equal to those from the original regression of $y$ on $X$ and $t$. Thus, it does not matter whether we

first detrend the series, and run the regression, or estimate the regression with trend included. Those are equvalent strategies, and since the latter is less burdensome, it may be preferred by the innately lazy researcher.

Another issue that may often arise in time series data of quarterly, monthly or higher frequency is **seasonality**. Some economic variables are provided in **seasonally adjusted** form. In databanks and statistical publications, the acronym SAAR (seasonally adjusted at annual rate) is often found. Other economic series are provided in their raw form, often labelled NSA, or not seasonally adjusted. Seasonal factors play an important role in many series. Naturally, they reflect the seasonal patterns in many commodities' measures: agricultural prices differ between harvest periods and out-of-season periods, fuel prices differ due to winter demand for oil and natural gas, or summer demand

for gasoline. But there are seasonal factors in many series we might consider with a more subtle interpretation. Retail sales, naturally, are very high in the holiday period: but so is the demand for cash, since shoppers and gift-givers will often need more cash at that time. Payrolls in the construction industry will exhibit seasonal patterns, as construction falls off in cold climates, but may be stimulated by a mild winter. Many financial series will reflect the adjustments made by financial firms to "dress up" quarter-end balance sheets and improve apparent performance.

If all of the data series we are using in a model have been seasonally adjusted by their producers, we may not be concerned about seasonality. But often we will want to use some NSA series, or be worried about the potential for seasonal effects. In this case, just as we dealt with trending series by including a time trend,

we should incorporate seasonality into the regression model by including a set of **seasonal dummies**. For quarterly data, we will need 3 dummies; for monthly data, 11 dummies; and so on. If we are using business-daily data such as financial time series, we may want to include "day-of-week" effects, with dummies for four of the five business days.

How would you use quarterly dummies in Stata? First of all, you must know what the time variable in the data set is: give the command `tsset` to find out. If it is a quarterly variable, the `tsset` range will report dates with embedded "`q`"s. Then you may create one quarterly dummy as `gen q1=(quarter(dofq(qtr))==1)` which will take on 1 in the first quarter, and 0 otherwise. To consider whether series `income` exhibits seasonality, `regress income L(1/3).q1` and examine the $F-$statistic. You could, of course, include any three of the four quarter dummies;

`L(0/2)` would include dummies for quarters 1, 2 and 3, and yield the same $F-$statistic. Note that inclusion of these three dummies will require the loss of at least two observations at the beginning of the sample. This form of seasonal adjustment will consider the effect of each season to be linear; if we wanted to consider multiplicative seasonality, e.g. sales are always 10% higher in the fourth quarter, that could be achieved by regressing $\log y$ on the seasonal dummies. A trend could be included in either form of the regression to capture trending behavior over and above seasonality; in the latter regression, of course, it would represent an exponential (constant growth) trend.

Just as with a trend, we may either deseasonalize each series (by regressing it on seasonal dummies, saving the residuals, and adding the mean of the original series) and regress seasonally adjusted series on each other; or we

may include a set of seasonal dummies (leaving one out) in a regression of $y$ on $X$, and test for the joint significance of the seasonal dummies. The coefficients on the $X$ variables will be identical, in both point and interval form, using either strategy.