Lecture 22

Time Series Regression: The Old-Fashioned Approach.

We are now going to use the subscript "t" in our multiple regressions instead of "i" to represent the fact that our attention is now turning to time series regression as opposed to cross-section regression.

Let y_t denote time-series observations on a dependent variable whose variation we would like to explain as a function of time-series observations on the k explanatory (independent) variables $x_1, x_2, \cdots, x_k$. When the $y, x_1, \cdots, x_k$ have trends in them but are, apart from trend, uncorrelated to each other, traditional multiple regression of the form

$$y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + u_t\,, \qquad (1)$$
$$t = 1, 2, \cdots, T.$$

are likely to indicate a statistically significant relationship when in fact there is none. (See for example the SAS program spurious.sas for an illustration of this point.) The <u>Old-Fashioned Approach</u> (prior to the 1980's) to avoiding this possible spurious inference was to use the following multiple regression

$$y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \gamma t + u_t \qquad (2)$$

where the variable t is a time-trend index variable that takes the values $t = 1, 2, \cdots, T$. Let the trend lines fitted to the variables $y, X_1, \cdots, X_k$ be denoted by

$$\begin{aligned} \tilde{y}_t &= a + bt \\ \tilde{X}_{t1} &= a_1 + b_1 t \\ &\vdots \\ \tilde{X}_{tk} &= a_k + b_k t \end{aligned} \qquad (3)$$

and the detrended values of y, $x_1, \cdots, x_k$ be denoted by

$$y_t - \tilde{y}_t = \ddot{y}_t$$

$$x_{t1} - \tilde{x}_{t1} = \ddot{x}_{t1} \qquad (4)$$

$$\vdots$$

$$x_{tk} - \tilde{x}_{tk} = \ddot{x}_{tk} .$$

Then by the Frisch-Waugh theorem, if we run the regression

$$\ddot{y}_t = \theta_0 + \theta_1 \ddot{x}_{t1} + \cdots + \theta_k \ddot{x}_{tk} + v_t \qquad (5)$$

and get the OLS estimates $\hat{\theta}_1, \cdots, \hat{\theta}_k$ (with $\hat{\theta}_0 \equiv 0$) it can be shown that these estimates are numerically identical to the OLS estimates that would result from applying OLS to equation (2), i.e. $\hat{\theta}_1 = \hat{\beta}_1; \cdots \hat{\theta}_k = \hat{\beta}_k$. Thus, the presence of t in equation (2) amounts to a trend control for the potentially trending variables y, $x_1, \cdots, x_k$ and thus reduces the possibility of deriving spurious correlation results. Having the trend term in (2) forces Ordinary Least Squares to examine the associations between the detrended values of $y, x_1, \ldots, x_k$, rather than the trending variables themselves.

Controlling for Seasonality

Not only do we need to control for trend in time series regressions by including a trend term (i.e. t) but also many time series observations exhibit seasonal patterns in that (if we are dealing with monthly data) a given month of each year tends to behave similarly relative to trend year after year. For exaple, retail sales of many companies tend to be above trend in December due to christmas sales while sales during January tend to be below trend following the christmas "splurge." In a similar manner sales in August and September tend to be above trend because of the beginning of the school year.

There are essentially three equivalent ways to model seasonality in time series multiple regression equations. Let assume that we are dealing with monthly retail sales of a company y_t, we have k variables to explain the variation in y_t (around trend) and let the monthly dummies be defined as

$$D_{tj} = \begin{cases} 1 & \text{if } t\text{-th observation is taken during the } j\text{-th month } (j=1,2,\cdots,12) \\ 0 & \text{otherwise.} \end{cases}$$

For example, ~~These~~ D_{t1} is the January monthly dummy which is 1 for all Januarys over which we observe the time series data and D_{t2} is the February monthly dummy, etc.

Now consider the following regression equation

$$y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \gamma \cdot t$$
$$+ \delta_1 D_{t1} + \delta_2 D_{t2} + \cdots + \delta_{12} D_{t12} + u_t . \quad (6)$$

Equation (6) is subject to the Dummy Variable Trap in that the monthly dummy variables $D_{t1}, \cdots, D_{t12}$ are exhaustive and when added up are perfectly collinear with the intercept term (a column of ones). ~~to~~ To get out of this trap (where OLS estimates are not computable) we need to either (i) drop one of the seasonal dummies making that month the base (reference) month or (ii) drop the intercept term ($\beta_0 = 0$) and run the regression through the origin or (iii) place a constraint on the dummy variable coefficients like $\delta_1 + \delta_2 + \cdots + \delta_k = 0$. These three different approaches give rise to three numerically equivalent models.

<u>Parametrization 1</u> (Relative to January)

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \gamma t$$

$$+ \delta_2 D_{t2} + \cdots + \delta_{12} D_{t2} + u_t \qquad (7)$$

where we have set $\delta_1 = 0$ and dropped D_{t1} from (6). The base trend line for January is therefore $y_t = \beta_0 + \gamma t$. The trend lines for the other months are given by $y_t = (\beta_0 + \delta_i) + \gamma t$, $i = 2, 3, \cdots, 12$. Then the δ_i, $i = 2, 3, \cdots, 12$ represent the differences of the intercepts for the months February through December relative to January. To test for the <u>absence</u> <u>of</u> <u>seasonality</u> in this model we would test the null hypothesis

$$H_0 : \delta_2 = \delta_3 = \cdots = \delta_{12} = 0. \qquad (8)$$

The F-statistic associated with this test would have 11 numerator degrees of freedom and $(N-k-1-11)$ denominator degrees of freedom.

Parametrization 2 (Each Month has its own intercept)

$$y_t = \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \gamma t$$
$$+ \delta_1 D_{t1} + \cdots + \delta_{12} D_{t12} + u_t . \qquad (9)$$

Here we have set $\beta_0 = 0$ in equation (6). The trend lines for each month are then given by $y_t = \delta_j D_{tj} + \gamma t = \delta_j + \gamma t$ for the months $j = 1, 2, \cdots, 12$. Each month's time trend has its own intercept, δ_j.

To test for the absence of seasonality in this model we test

$$H_0: \delta_1 = \delta_2 , \delta_2 = \delta_3 , \cdots , \delta_{11} = \delta_{12} . \qquad (10)$$

As before the F-statistic associated with this test would have 11 numerator degrees of freedom (due to the above 11 equality constraints) and $(N-k-12)$ denominator degrees of freedom.

The F-statistic calculated using parametrization 2 and the null hypothesis (10) will be numerically identical to the F-statistic calculated using parametrization 1 and the null hypothesis (8).

<u>Parametrization 3</u> (Seasonal Effects Sum to 0)

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \gamma t$$
$$+ \delta_1 D_{t1} + \cdots + \delta_{12} D_{t12} + u_t \qquad (11)$$

with $\delta_1 + \delta_2 + \cdots + \delta_{12} = 0$ imposed on model (11).

The Dummy Variable Trap is avoided because of the constraint $\delta_1 + \delta_2 + \cdots + \delta_{12} = 0$ is imposed.

The nice thing about this last parametrization is that the signs and magnitudes of the $\hat{\delta}_j$ coefficients helps us easily distinguish between the months that are "strong" relative to trend $(\hat{\delta}_j > 0)$ compared to the months ~~them~~ that are "weak" relative to trend $(\hat{\delta}_j < 0)$.

To test for the absence for seasonality in this model we test

$$H_0: \delta_2 = \delta_3 = \cdots = \delta_{12} = 0 \qquad (12)$$

For if $\delta_1 + \delta_2 + \cdots + \delta_{12}$ holds with (12) holding, we know that it also must be the case that $\delta_1 = 0$. As in the above parametrizations the identical F-statistic that we get here has 11 numerator degrees of freedom and $(N-k-12)$ denominator degrees of freedom.

To see an example of how these techniques are applied to a real economic times, sales of wood flooring by the members of the U.S. Wood Flooring Association see the SAS program wfa.sas and the documentation therein.