14.32 Econometric Data Science
Professor Anna Mikusheva
December, 2019

Lectures 24-26

Time Series

Notation

Time Series is a type of data collected on the same measure/unit over multiple periods of time. A clear characteristic of such data is that observations cannot be i.i.d, but rather are dependent.

Denote one observation as Y_t- observation at time t. Our data is $Y_1, ..., Y_T$, where time runs in an increasing direction. We assume that the data is sampled consecutively and equally-spaced.

Lag is an observation at the period just before, that is, Y_{t-1}. The j-th lag is Y_{t-j}. Difference is $\Delta Y_t = Y_t - Y_{t-1}$. Often the first difference of the log is called the growth rate:

$$\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1}) \approx \frac{\Delta Y_t}{Y_{t-1}}.$$

Often to report the growth rate we multiply it by 100. If data is reported quarterly, then to report a growth rate in annual terms we multiply it by 4. For example, imagine one has quarterly observations on the consumer price index (CPI). To report inflation in percents in annual rate we report

$$Inflation_t \approx 400\Delta \ln(CPI_t).$$

Stationarity Sequence $\{Y_t, t = ...-1, 0, 1, 2, ...\}$ is stationary if for any k the distribution of $\{Y_{t+1}, ...Y_{t+k}\}$ does not depend on t.

Auto-covariance $\gamma_k = cov(Y_t, Y_{t+k})$- for stationary series it depends only on the lag difference k, but not on t.

Forecasting

With time series data we often are faced with the task of forecasting (predicting the next observation). This task is very different from the causal effect estimation we've discussed in all previous lectures. In a forecasting regression: we cannot interpret coefficients, we do not think about omitted variable bias, we include regressors based on their predictive power.

Autoregression: AR(1)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + e_t.$$

Here β_0, β_1 do not have a causal interpretation. If $\beta_1 = 0$ then Y_{t-1} is not useful for forecasting Y_t. The coefficients can be estimated by OLS. Testing is handled in the usual way.

Terminology *Predicted value* is the value suggested by the regression for in-sample observations. *Forecasted value* is the value suggested by the regression for out-of-sample observations. We denote $Y_{T+1|T}$ to be the forecasted value for Y_{T+1} based on observations known at time T and a true population regression.

$$Y_{T+1|T} = \beta_0 + \beta_1 Y_T.$$

We denote $\widehat{Y}_{T+1|T}$ to be the forecasted value for Y_{T+1} based on observations known at time T and our estimated regression.

$$\widehat{Y}_{T+1|T} = \widehat{\beta}_0 + \widehat{\beta}_1 Y_T.$$

A multi-step forecast is when we iterate several steps into the future and, we denote it as $\widehat{Y}_{T+h|T}$. The forecast error is

$$Y_{T+1} - \widehat{Y}_{T+1|T} = e_{T+1} + (\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1)Y_T.$$

We think about e_{T+1} as unavoidable, but the error due to estimation will decline to zero with an increasing sample size. The root of the mean-squared forecast error characterizes a typical mistake:

$$MSFE = E(Y_{T+1} - \widehat{Y}_{T+1|T})^2 = E(e_{T+1}^2) + E\left((\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1)Y_T\right)^2.$$

The sum is due to the fact that future error e_{T+1} is uncorrelated to the estimation. Term $E(e_{T+1}^2)$ can be estimated from the SER (standard error of regression), while the second term will come from variances of the OLS estimates.

Another way of getting MSFE is via an out-of sample forecast: simulating the prediction exercise as if in real time.

It is common to report uncertainty about the forecast by reporting 'forecast intervals': $\widehat{Y}_{T+1|T} \pm 1.96\sqrt{MSFE}$. Unfortunately, the justification for such an interval is somewhat shaky - it relies on the assumption that e_{T+1} has a Gaussian distribution, which does not have to be true.

Autoregression: AR(p)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + e_t.$$

The coefficients do not have a causal interpretation. They can be estimated by OLS. Testing happens in the usual way: one coefficient can be tested by a t-test, multiple coefficients can be tested using an F-test.

Lag length selection. There is always a question of what lags to include (order of AR). The more recent lags are probably more informative than those more distant, so, there is an ordering. If one is deciding between $AR(p)$ and $AR(p-1)$, one can estimate $AR(p)$ and test $H_0 : \beta_p = 0$. However, choosing the order of AR by testing is not advisable, as it faces the problem of multiple testing. A better approach is to make use of the *Information Criterion*. For each p, run OLS to estimate coefficients in $AR(p)$ and calculate the sum of the squared residuals:

$$SSR(p) = \sum_t (y_t - \widehat{\beta}_0 - ... - \widehat{\beta}_p y_{t-p})^2.$$

As we discussed before, $SSR(p)$ is always decreasing in p. One would calculate BIC:

$$BIC(p) = ln\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{ln(T)}{T},$$

and choose $\widehat{p} = \arg\min_p BIC(p)$. The second term in BIC is the 'penalty' for estimating too many coefficients, which is intended to avoid over-fitting.

Autoregressive distributed lag model (ADL). One may use some other variables helpful in forecasting, call them X_t:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + \delta_1 X_{t-1} + ... + \delta_r X_{t-r} + e_t.$$

is an ADL(p, r) model. Again, if the goal is only to forecast, then the variables are added as long as they are useful for a forecast, and the coefficients do not have a causal interpretation. If one wants to test whether X has 'predictive power' for Y (it is referred to as a 'Granger causality') one should use an F-statistic for the hypothesis $H_0 : \delta_1 = ... = \delta_r = 0$. The order of the model p and r can be determined by BIC as well when one uses $1 + p + r$ in place of $p + 1$ in the penalty term.

Dynamic Causal Effect

The definition of causal effect is essentially the same through out the mental experiment when comparing potential outcomes for different values of treatment. 'Dynamic' stays for the possibility that a change in X may affect Y for several periods of time.

We consider a setting when we want to estimate the dynamic treatment effect from a long time series of (X_t, Y_t) rather than a cross-section. The concept of the idealized randomized control trial changes, as we do not have many entities/individuals to whom we can randomly assign treatment. Rather, we have only one entity observed for a long time. One can consider that we may take this long history and sub-divide it into many shorter ones. If two pieces of history are distant enough from each other, they will be almost independent and can be considered as different 'entities' in this random experiment. Thus, our randomized control trial will consist in randomly assigning different levels of treatment to different periods. Then we would estimate the following regression:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + e_t.$$

- β_0 is the impact effect (instantaneous or same period effect on Y_t from a change in X_t, while holding past $X_{t-1}, ..., X_{t-p}$ constant).
- β_1 is a 1-period dynamic multiplier (effect on Y_t from a change in X_{t-1}, while holding $X_t, X_{t-2}, ..., X_{t-p}$ constant).
- β_2 is 2-period dynamic multiplier (effect on Y_t from a change in X_{t-2} while holding $X_t, X_{t-1}, X_{t-3}, ..., X_{t-p}$ constant).
- $\beta_0 + \beta_1 + ... + \beta_p$ is a cumulative dynamic multiplier.

Assumptions for OLS:

Assumption 1 Exogeneity: $E(e_t|X_t, X_{t-1}, ...\text{all past values}) = 0$.

Assumption 2 (a) (Y_t, X_t) is a stationary process. (b) (Y_t, X_t) and (Y_{t-j}, X_{t-j}) become close to independent as j grows. (Technical assumption is quite complicated).

Assumption 3 No outliers: $Ee_t^8 < \infty$, $EX_t^8 < \infty$.

Assumption 4 No perfect multicollinearity.

There are a Law of Large Numbers and Central Limit Theorems for time series data, but they require more technically involved assumptions (thus Assumptions 2 and 3).

Under Assumptions 1-4 OLS is consistent and asymptotically gaussian. There is one challenge though – related to standard errors and the fact that time series data tend to be auto-correlated.

Heteroskedasticity and autocorrelation robust (HAR) standard errors. Consider a simpler regression using time series data:

$$Y_t = \alpha + \beta X_t + e_t.$$

By the usual arguments

$$\sqrt{T}(\widehat{\beta} - \beta) = \frac{\frac{1}{\sqrt{T}} \sum_t (X_t - \bar{X}) e_t}{\frac{1}{T} \sum_t (X_t - \bar{X})^2}.$$

Also as usual, we have Law of Large Numbers $\frac{1}{T} \sum_t (X_t - \bar{X})^2 \to Var(X)$. The non-standard part is the numerator

$$\frac{1}{\sqrt{T}} \sum_t (X_t - \bar{X}) e_t = \frac{1}{\sqrt{T}} \sum_t \xi_t.$$

There is a Central Limit Theorem applicable to this case, but the main difference lies in the variance. If ξ_i is an i.i.d. sequence then

$$Var\left(\frac{1}{\sqrt{T}}\sum_t \xi_t\right) = \frac{1}{T}Var\left(\sum_t \xi_t\right) = \frac{1}{T}\sum_t Var\left(\xi_t\right) = Var(\xi).$$

If the random variables are correlated then this formula does not hold any more. For example if $T = 2$:

$$Var\left(\frac{\xi_1 + \xi_2}{\sqrt{2}}\right) = \frac{1}{2}\left(Var(\xi_1) + Var(\xi_2) + 2cov(\xi_1, \xi_2)\right) = \gamma_0 + \gamma_1,$$

where $Var(\xi_1) = Var(\xi_2) = \gamma_0$ and $cov(\xi_1, \xi_2) = \gamma_1$ from stationarity. One can show that for arbitrary T the variance formula depends on multiple autocovariances:

$$Var\left(\frac{1}{\sqrt{T}}\sum_t \xi_t\right) = \gamma_0 + 2\sum_{k=1}^{T}\left(\frac{T-k}{T}\right)\gamma_k.$$

A consistent estimator in this case is known as HAR(HAC) or Newey-West. It estimates several autocovariances (though not all available) and uses them in somewhat similar formula. There is a cut-off parameter m that is a researcher's choice. A rule of thumb is $m = 0.75 \cdot T^{1/3}$.