ØAMET4100 · Spring 2019
Lecture Note 9

Instructor: Fenella Carpena

March 21, 2019

This lecture note provides a review of time series data (Stock & Watson, Chapter 14). This lecture note is not intended to be a comprehensive review of lecture or the textbook, since there is a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Basic Concepts and Terminology

In this lecture, we apply regression analysis to forecasting. The empirical example we will use throughout is the following: How fast will **Gross Domestic Product (GDP)** grow in the next quarter? Before we begin, we need to learn some basic concepts and terminology about time series.

Time series data is data for the same entity over multiple time periods. Suppose our time series data is **quarterly GDP of the US (measured in billions of dollars), 1960-2012**. If we were to look at this time series data like a "spreadsheet" in Microsoft Excel, it would look like this:

Time (t)	GDP (Y_t)	First Lag (Y_{t-1})	Third Lag (Y_{t-3})	First Difference (ΔY_t)
1960:Q1	3120	.	.	.
1960:Q2	3108	3120	.	−12
1960:Q3	3116	3108	.	8
1960:Q4	3078	3116	3120	−38
1961:Q1	3099	3078	3108	21

A particular observation Y_t is indexed by the subscript t, where t denotes the time period (e.g., quarters). The total number of observations in a time series is T (i.e., the number of periods).

With time series data, we have special terminology and notation for future and past values of Y.

- Y_{t-1}: This is called the **first lagged value** (or first lag). If t is the current period, Y_{t-1} is the value of Y in the previous period.
- Y_{t-j}: This is the j^{th} lag.
- Y_{t+j}: This is the j^{th} future value.
- ΔY_t: This is called the **first difference**, and is equal to $Y_t - Y_{t-1}$. This is the change in Y between periods t and $t-1$.

Time series data are often analyzed with variables in logarithmic form. This is because many economic time series (e.g., GDP) exhibit growth that is approximately exponential. For example, in the long run, GDP grows by a given percentage per year on average. Thus, the logarithm of the GDP grows approximately linearly.

- $\Delta ln(Y_t) = ln(Y_t) - ln(Y_{t-1})$, is the first difference in the logarithm of Y_t. Note that $ln(Y_t) - ln(Y_{t-1}) = ln(Y_t/Y_{t-1}) = ln(1 + \frac{Y_t - Y_{t-1}}{Y_{t-1}}) \approx \frac{\Delta Y_t}{Y_{t-1}}$ when $\frac{\Delta Y_t}{Y_{t-1}}$ is small (see equation 8.16 of the textbook for more details).
- $100 * \Delta ln(Y_t)$ is the percentage change of a time series Y_t between periods $t-1$ and t. This approximation is most accurate when the percentage change is small.

In our example with GDP, we have data on quarters, so if we wanted to annualize this percentage change, we would multiple by 4 to obtain $4 * 100 * \Delta ln(Y_t)$.

Time (t)	GDP (Y_t)	$ln(GDP_t)$	$\Delta ln(GDP_t)$	Annual GDP Growth Rate ($GDPGR_t$)
1960:Q1	3120	8.046	.	.
1960:Q2	3108	8.042	−0.004	−1.6%
1960:Q3	3116	8.044	0.002	0.8%
1960:Q4	3078	8.032	−0.012	−4.8%
1961:Q1	3099	8.039	0.007	2.8%

In time series data, the value of Y in one period is typically correlated with its value in the next period. Thus, Y is said to be **autocorrelated** or **serially correlated**.

- The j^{th} **autocovariance** of a series Y_t is the (population) covariance between Y_t and its j^{th} lag. We write it as $cov(Y_t, Y_{t-j})$. It is estimated using the sample j^{th} autocovariance

$$\widehat{cov(Y_t, Y_{t-j})} = \frac{1}{T}\sum_{t=j+1}^{T}(Y_t - \overline{Y}_{j+1:T})(Y_{t-j} - \overline{Y}_{1:T-j})$$

  where $\overline{Y}_{j+1:T}$ denotes the sample average computed using the periods $j+1$ to T.

- The j^{th} **autocorrelation** of a series Y_t is the (population) correlation between Y_t and Y_{t-j}. We write it as $\rho_j = corr(Y_t, Y_{t-j})$. It is estimated using the the sample j^{th} autocorrelation

$$\widehat{\rho}_j = \frac{\widehat{cov(Y_t, Y_{t-j})}}{\widehat{var(Y_t)}}$$

  where the denominator refers to the sample covariance of Y. Note that the denominator in this formula assumes that $var(Y_t) = var(Y_{t-j})$, which is an implication of the assumption that Y is **stationary**. Stationarity is a very important concept in time series that we will delve into in more detail next lecture.

2 Autoregressions: AR(1) Model

In this section, we will look at forecasting using **autoregression**, a regression model that relates a time series variable to its past values. We abbreviate autoregression as "AR."

Why do we care about AR models? We care because if you want to forecast the future of a time series (e.g., the rate of GDP growth next quarter), a good place to start is to use information from the immediate past (e.g., how fast GDP grew last quarter).

2.1 Model Specification

For the time series Y_t, the population regression for a **first-order autoregression** or **AR(1)** model is written as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t \tag{1}$$

where $E(u_t|Y_{t-1}, \ldots) = 0$. As can be seen in the above equation, a first-order autoregression or AR(1) model forecasts Y by using the previous period's value. We call this an "autoregression" because it is a regression of the series onto its own lag (in ancient Greek, the word "auto" means "self"). The regression is "first-order" because only one lag is used in the regression.

2.2 Estimation

Using time series data with periods $t = 1, \ldots, T$, we can estimate equation (1) using OLS. Importantly, note that we only have T periods in our data, so forecasting means we want to predict the value of Y for period $T+1$, given information that we have until period T.

2.3 Forecasts and Forecast Errors

If Y_t follows an AR(1) process and β_0 and β_1 are known, then our forecast for Y_{T+1}, based on what we know at the current period Y_T, is $\beta_0 + \beta_1 Y_T$. However, the population β's are unknown, so we use the OLS estimates $\widehat{\beta}$'s. Specifically, denote $\widehat{Y}_{T+1|T}$ as the forecast for Y_{T+1} based on information through period T.

$$\widehat{Y}_{T+1|T} = \widehat{\beta}_0 + \widehat{\beta}_1 Y_T \tag{2}$$

The **forecast error** is then the mistake made by the forecast: it is the difference between our forecasted value $\widehat{Y}_{T+1|T}$ and the value that actually occurred Y_{T+1}.

$$\text{Forecast Error} = Y_{T+1} - \widehat{Y}_{T+1|T} \tag{3}$$

Note that the forecast $\widehat{Y}_{T+1|T}$ is made at time T, and we can only calculate the forecast error once we've reached (or passed) time $T+1$, i.e., after we find out the actual value of Y_{T+1}.

The forecast $\widehat{Y}_{T+1|T}$ is **NOT** the same as the OLS predicted value, and the forecast error is **NOT** the same as the OLS residual. Why? OLS predicted values $\widehat{Y}_t$ and residuals $\widehat{u}_t = Y_t - \widehat{Y}_t$ for $t \leq T$ are "in-sample." This means that they are both calculated for the observations in our sample, which is what we used to estimate the regression. In contrast, the forecast $\widehat{Y}_{T+1|T}$ and the forecast error $Y_{T+1} - \widehat{Y}_{T+1|T}$ are "out-of-sample." This means they are calculated for some date beyond what we have in the data. Because our data has only T periods, Y_{T+1} is not observed in the data we used to estimate the regression.

2.4 Mean Squared Forecast Error (MSFE)

Because forecasts are uncertain, it is useful to have a measure of the uncertainty of our forecast. One measure is the MSFE, which is the spread of the forecast error distribution.

$$\text{MSFE} = E[(Y_{T+1} - \widehat{Y}_{T+1|T})^2] \tag{4}$$

The Root Mean Squared Forecast Error (RMSFE) is the square root of the MSFE, and it gives us the magnitude of a typical mistake made using a forecasting model.

$$\text{RMSFE} = \sqrt{E[(Y_{T+1} - \widehat{Y}_{T+1|T})^2]} \tag{5}$$

What is the value of $E[(Y_{T+1} - \widehat{Y}_{T+1|T})^2]$? For simplicity, we assume u_t is homoskedastic. In the AR(1) model, $Y_{T+1} - \widehat{Y}_{T+1|T} = \beta_0 + \beta_1 Y_t + u_{t+1} - \widehat{\beta}_0 - \widehat{\beta}_1 Y_T = u_{t+1} - [(\widehat{\beta}_0 - \beta_0) + (\widehat{\beta}_1 - \beta_1) Y_T]$. Thus,

$$\text{MSFE} = \sigma_u^2 + var[(\widehat{\beta}_0 - \beta_0) + (\widehat{\beta}_1 - \beta_1) Y_T] \tag{6}$$

We can see from the expression in equation (6) that the MSFE has two sources of error.

1. The error arising because future values of u_t are unknown, σ_u^2.
2. The error in estimating the regression coefficients, $var[(\widehat{\beta}_0 - \beta_0) + (\widehat{\beta}_1 - \beta_1) Y_T]$.

The first term, σ_u^2, can be estimated by the square of the standard error of the regression (SER). The second term in equation (6) requires estimating the variance of a weighted average of regression coefficients, and one approach is to use the variance of pseudo out-of-sample forecasts, which we will discuss in next week's lecture.

Note that if the sample size is large (and our regression specification is correct), we get consistent estimates of the regression coefficients. A large sample size thus implies that the errors from the first source are larger than in the second source. As a result, RMSFE $\approx \sigma_u$, and as mentioned above, σ_u can be estimated by the standard error of the regression, $SER = \frac{1}{T-2} \sum_{t=1}^{T} \widehat{u}_t^2$.

2.5 Forecast Intervals

A forecast interval is like a confidence interval except that it pertains to a forecast. For example, a 95% **forecast interval** is an interval that contains the future value of the series in 95% of repeated applications.

The 95% forecast interval is given by $\widehat{Y}_{T+1|T} \pm 1.96SE(Y_{T+1} - \widehat{Y}_{T+1|T})$, where $SE(Y_{T+1} - \widehat{Y}_{T+1|T})$ is an estimator of the RMSFE. Note that this forecast interval is calculated under the assumption that the forecast error is normally distributed. This assumption follows if u_{T+1} is normally distributed, which implies that the central limit theorem applies to the regression coefficients. As a result, the forecast error is the sum of two independent, normally distributed terms (i.e., Y_{T+1} and $\widehat{Y}_{T+1|T}$), and the forecast error itself is normally distributed.

2.6 Example: Forecasting GDP Growth Rates

Suppose that our Y variable corresponds to $GDPGR$ (i.e., the growth rate of GDP), and we have a quarterly time series of $GDPGR$ from 1962:Q1 to 2012:Q4 (i.e., our data stops in 2014, Quarter 4). Using these data, we can estimate equation (1) using OLS. See the Stata do-file for this lecture.

The OLS regressions result is

$$\widehat{GDPGR}_t = \underset{(0.35)}{1.99} + \underset{(0.08)}{0.34}\, GDPGR_{t-1},\ \overline{R}^2 = 0.1,\ SER = 3.16,\ N = 204 \tag{7}$$

The following are the values of $GDPGR$ in from 2012:Q1 to 2013:Q1.

Date	2012:Q1	2012:Q2	2012:Q3	2012:Q4	2013:Q1
$GDPGR$	3.64	1.20	2.75	0.15	1.1

Is a positive growth rate of GDP in one quarter associated with positive growth in the next quarter? Yes, we find that the coefficient $\widehat{\beta}_1$ is greater than zero. Therefore, positive growth of GDP in one quarter is associated with positive growth in the next quarter.

What is the forecast for the growth rate of GDP in 2013:Q1 that we would have made in 2012:Q4, based on the above AR(1) model? The growth rate in 2012:Q4 is 0.15%, so our forecast using the AR(1) model is $1.99 + 0.34 * 0.15 = 2.0$.

Is this forecast accurate? What is the forecast error? As it turns out, the actual growth rate in 2013:Q1 was 1.1%, so the forecast error is -0.9 percentage points. From the above results, we see that the adjusted R^2 is only 0.1, which means that the lagged value of $GDPGR$ explains only a small fraction of the variation in GDP growth in our sample data. This low adjusted R^2 is consistent with our poor forecast for 2013:Q1.

Ignoring the uncertainty arising from the estimation of the coefficients, what is the estimate of the RMSFE? Explain in words what this estimate means. The SER of our OLS regression is 3.16. If we ignore the uncertainty arising from the estimation of the coefficients, our estimate of the RMSFE is therefore, 3.16 percentage points. This means that when we use our AR(1) model, the typical magnitude of the mistake in the GDP growth rate forecast is 3.16 percentage points.

3 Autoregressions: AR(p) model

3.1 Model Specification

By using only the first lag, the AR(1) model ignores potentially useful information from much earlier periods. Using p lags yields the p^{th}**-order autoregressive** or **AR(p)** model. The AR(p) model is written as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + u_t \tag{8}$$

where $E(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$. The number of lags p is called the "order" (or the lag length) of the autoregression.

The assumption $E(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$ means that the conditional expectation of u_t is zero, given all past values if Y. This assumption has two important implications.

1. The best forecast of Y_{T+1} based on its entire history depends on only its most recent p past values. Why is this the case? Note that $E(u_t|Y_{t-1}, Y_{t-2}, \ldots) = 0$ implies that $E(Y_t|Y_{t-1}, Y_{t-2}, \ldots) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p}$; this means that the conditional expectation function of Y_t corresponds exactly to the population regression line. One can then show that for any t, the forecast $Y_{t|t-1}$ for Y_t that minimizes the conditional RMSFE is equal to $E(Y_t|Y_{t-1}, Y_{t-2}, \ldots)$.
2. The errors u_t are serially uncorrelated.

3.2 Lag Length Selection Using Information Criteria

How many lags should be included in the regression? In other words, what order p should we use? In practice, selecting the "right" p requires finding the "right" balance. If p is too low, we might be throwing away potentially useful information from lags that are further away. If p is too high, we will be estimating more coefficients than necessary, which introduces more estimation error in our forecasts.

F-statistic Approach. The idea behind this approach is to start with a model with many lags, and then perform a hypothesis test on the final lag. For example, estimate an AR(6) and test if the coefficient on the sixth lag is significant at the 5% level. If not, drop the sixth lag and estimate an AR(5), then test the fifth lag, etc.

The disadvantage of this approach is that it can produce a model that is too large some of the time. To see this, suppose we conduct our tests at a 5% significance level. If the true $p = 5$, and because our Type I error rate is 5%, we will (incorrectly) conclude that $p = 6$ just by chance 5% of the time.

Bayes Information Criterion (BIC). The Bayes Information Criterion, also called the Schwarz Information Criterion (SIC), is given by

$$BIC(p) = ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{ln(T)}{T} \tag{9}$$

where $SSR(p)$ is the sum of squared residuals of the estimated AR(p) model, $p + 1$ is the number of coefficients we estimate in the model (including the regression constant), and T is the number of time periods in the data. What is the intuition behind this formula? The first term, which contains $SSR(p)$, is weakly decreasing when we add more lags. The second term, which contains $p+1$, increases as we add more lags. Thus, the BIC trades of these two forces, so that the number of lags is a consistent estimator of the true lag length.

How do we select the lag length p using the BIC? We calculate the value of BIC for every candidate model, and the "best" model is the one with the smallest BIC.

1. Set out the possible choices for the lag length $p = 0, 1, \ldots, p_{max}$, where p_{max} is the largest value of the lag length that we will consider. So our candidate models are AR(0), AR(1), AR(2), ..., AR(p_{max}).
2. Calculate the BIC for every candidate model.
3. Find $\widehat{p}$, which represents the BIC estimator of p. Specifically, $\widehat{p}$ is the value of p that has the smallest BIC out of all the candidate models.

Akaike Information Criterion (AIC). Another information criterion is due to Akaike (1974). The AIC is given by

$$AIC(p) = ln\left[\frac{SSR(p)}{T}\right] + (p+1)\frac{2}{T} \tag{10}$$

Note that the BIC and AIC are quite similar, except that the term $\frac{ln(T)}{T}$ in the BIC is replaced with $\frac{2}{T}$. What differences does this make between the BIC and the AIC?

1. When T is large, $ln(T) > 2$. Thus, a smaller decrease in SSR is needed in the AIC to justify including another lag. Another way of saying this is that the BIC imposes a greater "penalty" for adding lags.

2. The AIC estimator of p is not consistent, but is generally more efficient that the BIC. This means that the AIC and BIC have their own advantages and disadvantages. When the sample size is very large, the BIC will give us the correct model, and the AIC generally overestimates p. However, the average variation in the selected model orders from different samples within a given population will be greater in the context of BIC than AIC.

Overall, then, no criterion is definitely superior to others, and this is an important topic of debate and research in the statistics literature. Note that when using BIC or AIC, all candidate models should be estimated using the same set of observations (i.e., they should all have the same sample size).

3.3 Example: Forecasting GDP Growth Rates

Going back to our example with GDP growth rates, we can estimate an AR(2) model using OLS. See the Stata do-file. The OLS regression results is

$$\widehat{GDPGR}_t = \underset{(0.40)}{1.63} + \underset{(0.08)}{0.28}\, GDPGR_{t-1} + \underset{(0.08)}{0.18}\, GDPGR_{t-2},\ \overline{R}^2 = 0.14,\ SER = 3.11,\ N = 204 \quad (11)$$

We also estimate an AR(0) model (i.e, corresponding to a regression on a constant), and the OLS regression result is

$$\widehat{GDPGR}_t = \underset{(0.23)}{3.05},\ \overline{R}^2 = 0.00,\ SER = 3.35,\ N = 204 \quad (12)$$

What is the forecast for the growth rate of GDP in 2013:Q1 that we would have made in 2012:Q4, based on the above AR(2) model? What is the forecast error? To compute the forecast, we just need to plug-in the values of $GDPPR$ in 2012:Q4 in the first lag and 2012:Q3 in the second lag. Doing so, we get $1.63 + 0.28 * 0.15 + 0.18 * 2.75 = 2.1\%$. Since the actual growth rate in 2013:Q1 is 1.1%, this means our forecast error is -1 percentage point.

How many lags should be included in the AR model? Suppose that our candidate models are AR(0), AR(1) and AR(2). Use the F-statistic method, BIC, and AIC to determine whether the AR model should be AR(0), AR(1), or AR(2). Let's start with the F-statistic method. In the AR(2) model, we can see that the coefficient on the second lag is statistically significant at the 5% level: the F-statistic (i.e., $(0.18/0.08)^2 \approx 5$) is greater than the critical value of 3.84; this finding suggests that the second lag belongs in the model.

Now, let's estimate the BIC and AIC. The calculations are shown in the table below, where I use the fact that $SER = \sqrt{\frac{SSR}{n-k-1}}$, where n is the sample size and k is the number of regressors.

	$p=0$	$p=1$	$p=2$
SER	3.35	3.16	3.11
SSR	2256	2017	1944
$ln(SSR(p)/T)$	2.40	2.29	2.25
$(p+1)*ln(T)/T$	0.03	0.05	0.08
$(p+1)*2/T$	0.01	0.02	0.03
BIC	2.43	2.34	2.33
AIC	2.41	2.31	2.28

As we can see in the above calculations, the BIC is minimized when $p = 2$. Similarly, the AIC is minimized when $p = 2$. These results therefore suggest that among the candidate models, the most appropriate model is AR(2).

4 Autoregressive Distributed Lag: One Additional Regressor

Economic theory often suggests other factors that could help forecast a variable of interest. When these other variables and their lags are added to an auto regression, we get an **autoregression distributed lag (ADL)** model.

4.1 Model Specification

An ADL model with p lags of the dependent variable Y_t and q lags of an additional predictor X_t, denoted **ADL(p,q)**, is written as

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + \delta_1 X_{t-1} + \ldots + \delta_q X_{t-q} + u_t \tag{13}$$

where u_t is the error term and the β's and the δ's are unknown coefficients. This model assumes that $E(u_t|Y_{t-1}, Y_{t-2}, \ldots, X_{t-1}, X_{t-2}, \ldots) = 0$, which implies that no lags of either Y or X belong in the ADL model. In other words, p and q are the true lag lengths, and the coefficients on the other lags are zero.

We call this model "autoregressive" lagged values of the dependent variable are included as regressors, and we call this model "distributed lag" because the regression also includes multiple lags (a "distributed lag") of an additional predictor.

4.2 Granger "Causality" Test

Do the included lags for X have useful predictive content for Y_t, beyond that contained in all other regressors? This claim can be tested using a two-sided null hypothesis test where $H_0 : \delta_1 = 0, \ldots, \delta_q = 0$. The F-statistic corresponding to this test is called the **Granger causality statistic**.

If the null hypothesis is rejected, we say that X Granger-causes Y. This **DOES NOT** mean that we have estimate the causal effect of X on Y! If X Granger-causes Y, this just means that X is a useful predictor for Y. Thus, "Granger predictability" is a more accurate term than "Granger causality," but unfortunately, the latter term has become part of the jargon of econometrics.

For example, we might be able to predict whether it will rain in the next hour by looking at how many people are carrying an umbrella during this hour. Thus, the number of umbrellas Granger-causes rainfall. But this does not mean that by making people carry umbrellas, we can cause rainfall.

4.3 Example: Forecasting GDP Growth Rates

We can augment the AR(2) model equation (11) by including the **term spread** as an additional variable in the autoregressive model. The term spread is the difference between long-term and short-term interest rates. The term spread generally falls before U.S. recessions, and this pattern suggests that the term spread might contain information about the future growth of GDP that is not already contained in past values of GDP growth. Adding the first and second lags of the term spread to equation (7) and using OLS, we obtain the following ADL(2,2) model.

$$\begin{aligned} \widehat{GDPGR}_t = & \underset{(0.48)}{0.97} + \underset{(0.08)}{0.24}\, GDPGR_{t-1} + \underset{(0.08)}{0.18}\, GDPGR_{t-2} \\ & - \underset{(0.43)}{0.14}\, TSpread_{t-1} + \underset{(0.44)}{0.66}\, TSpread_{t-2},\ \overline{R}^2 = 0.17,\ SER = 3.06,\ N = 204 \end{aligned} \tag{14}$$

An F-test for the joint significant of the coefficients on $TSpread$ yields an F-stat of 4.33. The following are the values of $TSpread$ from 2012:Q1 to 2013:Q1.

Date	2012:Q1	2012:Q2	2012:Q3	2012:Q4	2013:Q1
$TSpread$	1.97	1.74	1.54	1.62	1.86

Use the Granger causality statistic to determine if the term is a useful predictor for the GDP growth rate. The F-statistic on all the term spread coefficients is 4.33. Assuming a test at the 5% level, the critical value is 3.00. Therefore, we reject the null hypothesis that the coefficients on term spread are jointly zero at the 5% level.

Do the results from the above hypothesis test mean that a change in the term spread will cause a subsequent change in the GDP growth rate? No. Granger causality only means that the past value of the term spread appears to contain information that is useful for forecasting changes in GDP, beyond information that is already contained in lagged GDP growth rates.

What is the forecast for the growth rate of GDP in 2013:Q1 that we would have made in 2012:Q4, based on the above ADL(2,1) model? What is the forecast error? The forecast is the actual growth rate in 2013:Q1 is $0.97 + 0.24 * 0.15 + 0.18 * 2.75 - 0.14 * 1.62 + 0.66 * 1.54 = 2.3$. Since the actual growth rate in 2013:Q1 is 1.1%, this means our forecast error is -1.2 percentage point.

5 Autoregressive Distributed Lag: Multiple Predictors

We now extend the ADL model to include multiple predictors and lags. Note that because we have multiple predictor (e.g., term spread, inflation rate, unemployment rate, etc.) and their lags, we have two subscripts for the X variables and their coefficients: the first subscript indexes the predictor, and the second subscript indexes time. For example, $X_{2,t-1}$ refers to the first lag of the second predictor.

5.1 Model Specification

The general time series regression allows for k additional predictors, where q_1 lags of the first predictor are included, q_2 lags of the second predictor are included, etc. The regression model is written as:

$$\begin{aligned} Y_t = \beta_0 &+ \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} \\ &+ \delta_{1,1} X_{1,t-1} + \ldots + \delta_{1,q_1} X_{1,t-q_1} \\ &+ \ldots \\ &+ \delta_{k,1} X_{k,t-1} + \ldots + \delta_{k,q_k} X_{1,t-q_k} \end{aligned} \tag{15}$$

The model makes the following four assumptions.

1, $E(u_t | Y_{t-1}, Y_{t-2}, \ldots, X_{1,t-1}, X_{1,t-2}, \ldots, X_{k,t-1}, X_{k,t-2}, \ldots) = 0$. This means that u_t has conditional mean zero, given all the regressors **AND** all the lags of the regressors, even those not included in the model. This assumption implies that the best forecast of Y_t using all past values of Y and the X's is given by equation (15).

2a. The random variables $(Y_t, X_{1,t}, \ldots, X_{k,t})$ have a stationary distribution. This is the time series equivalent of the "identically distributed" part of the i.i.d. assumption that we have seen before.

Stationary means that the probability distribution does not change over time. Hence, the joint distribution of the random variables $(Y_t, X_{1,t}, \ldots, X_{k,t})$ when $t = 1$ is the same as when $t = 2$, etc. In words, this means that the distribution of the random variable (e.g., mean, variance, etc.) in the first period is the same as its distribution in the next period, etc. Intuitively, stationarity requires that the future is like the past (in a probabilistic sense).

Why do we care about stationarity? If the time series variables are not stationary, then the historical relationships of the variables in the past—which is what we are examining when we estimate an AR or ADL model—may not be reliable guides for the future. One or more problems can arise: the forecast can be biased, it can be inefficient, or conventional OLS-based statistical inferences (e.g., comparing the OLS t-stat with ± 1.96) can be misleading.

Stationarity is a very important concept in time series analysis, and as we will see later on, if a series is not stationary, then conventional hypothesis tests, confidence intervals and forecasts can be unreliable.

2b. $(Y_t, X_{1t}, \ldots, X_{kt})$ and $(Y_{t-j}, X_{1,t-j}, \ldots, X_{k,t-j})$ become independent as j gets large. This assumption is the time series equivalent of the "independent" part of the i.i.d. assumption that we have seen before. In words, this assumption means that the random variables become independently distributed as the time difference between them gets large. This assumption is also sometimes referred to as **weak dependence**.

3. Large outliers are unlikely: $X_{1,t}, \ldots, X_{k,t}$ and Y_t have nonzero, finite fourth moments.

4. There is no perfect multicollinearity.

Why do we care about these assumptions? We care because if they hold, we can proceed with out our usual approaches for statistical inference on the regression coefficients (e.g., F-statistics).

5.2 Lag Length Selection Using Information Criteria

To determine the number of lags (i.e., the values for p, q_1, ..., q_k), we can use the same approaches discussed in the last lecture.

F-statistic approach. As in the AR(p) models, we can use the F-statistic to test whether a coefficient is equal to zero.

BIC and AIC. If the regression model has K coefficients (including the regression constant β_0), the BIC is given by

$$BIC(p) = ln\left[\frac{SSR(p)}{T}\right] + K\frac{ln(T)}{T}. \tag{16}$$

The AIC is also defined in the same way, with 2 replacing $ln(T)$ in the second term. There are two important points to remember when using the BIC or the AIC in practice.

1. As with using the BIC and AIC in AR(p), all candidate models should be estimated using the same sample (i.e., the number of observations T in each candidate regression must be the same).
2. When we have multiple X's, using BIC and AIC is computationally demanding because there are many different combinations of lag parameters. So in practice, a convenient shortcut is to require all regressors to have the same number of lags. This means setting $p = q_1 = q_2 = \ldots = q_k$, so that only $p_{max} + 1$ models need to be compared.