ECON 7310 Elements of Econometrics
Week 11: Regression Model for Time-Series

David Du[1]

[1] University of Queensland

Draft

Time Series Data

Time series data are data collected on the same observational unit at multiple time periods.

- GDP for a country (e.g., 20 years of quarterly data).
- AUD/USD exchange rates (daily data for 1 year = 365 observations).
- Cigarette consumption per capita in California, by year (annual data).

Some uses of time series data

1. Forecasting (SW Chapter 15).
2. Estimation of dynamic causal effects (SW Chapter 16).
3. Modeling risks in financial markets (SW Chapter 17).
4. Applications outside of economics: environmental and climate modeling, engineering (system dynamics), CS (network dynamics)...

This class briefly introduces 1. ECON7350 is designed for a systematic learning of various time series models.

Notation

- Y_t = value of Y in period t.
- Data set: $\{Y_1, ..., Y_T\}$ are T observations on the time series variable Y.
- We consider only consecutive, evenly-spaced observations (e.g., monthly, 1960-1999, no missing months).

Lags, First Differences, and Growth Rates

- The first lag of a time series Y_t is Y_{t-1}; its jth lag is Y_{t-j}.
- The first difference of a series, ΔY_t, is its change between periods $t-1$ and t, i.e., $\Delta Y_t = Y_t - Y_{t-1}$.
- The first difference of the logarithm of Y_t: $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$.
- The percentage change of Y_t between periods $t-1$ and $t \approx 100 \times \Delta \ln(Y_t)$, where the approximation is most accurate when the percentage change is small.

Example: Quarterly Rate of Growth of U.S. GDP at an Annual Rate

- GDP = Real GDP in the US (Billions of \$2009).
- GDP in the fourth quarter of 2016 (2016:Q4) = 16851.
- GDP in the first quarter of 2017 (2017:Q1) = 16903.
- Percentage change in GDP, 2016:Q4 to 2017:Q1 is

$$100 \times \frac{16903 - 16851}{16851} = 0.31\%$$

- Percentage change in GDP, 2016:Q4 to 2017:Q1, *at an annual rate* $= 4 \times 0.31\% = 1.24\%$ (percent per year).
- Using the logarithmic approximation to percent changes yields $4 \times 100 \times [\ln(16903) - \ln(16851)] = 1.232\%$.

Autocorrelation

The correlation of a series with its own lagged values is called *autocorrelation* or *serial correlation*.

- The first *autocovariance* of Y_t is $Cov(Y_t, Y_{t-1})$.
- The first *autocorrelation* of Y_t is

$$Corr(Y_t, Y_{t-1}) = \frac{Cov(Y_t, Y_{t-1})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-1})}} = \rho_1$$

These are population correlations - they describe the population joint distribution of (Y_t, Y_{t-1}).

More generally...

Autocorrelation

Autocorrelation

The jth autocovariance of a series Y_t is the covariance between Y_t and its jth lag, Y_{t-j}, and the jth autocorrelation coefficient is the correlation between Y_t and Y_{t-j}. That is,

- The jth *autocovariance* of Y_t is $Cov(Y_t, Y_{t-j})$.
- The jth *autocorrelation* of Y_t is

$$Corr(Y_t, Y_{t-j}) = \frac{Cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-j})}} = \rho_j$$

The jth autocorrelation coefficient is sometimes called the jth serial correlation coefficient.

Sample Autocorrelation

The jth *sample autocorrelation* is an estimate of the jth population autocorrelation:

$$\hat{\rho}_j = \frac{\widehat{Cov}(Y_t, Y_{t-j})}{\widehat{Var}(Y_t)}$$

where $\widehat{Cov}(Y_t, Y_{t-j}) = T^{-1} \sum_{t=j+1}^{T} (Y_t - \bar{Y}_{j+1,T})(Y_{t-j} - \bar{Y}_{1,T-j})$.

- $\bar{Y}_{j+1,T}$ is the sample average of Y_t computed over observations $t = j + 1, ..., T$.

Note that

- the summation is over $t = j + 1$ to T. (why?)
- The divisor is T, not $T - j$ (this is the conventional definition used for time series data).

Forecasting

Forecasting and estimation of causal effects are quite different objectives.

For estimation of causal effects, we were very concerned about omitted variable bias, control variables, and exogeneity.

For forecasting,

- Omitted variable bias isn't a problem!
- We won't worry about interpreting coefficients in forecasting models – no need to estimate causal effects if all you want to do is forecast!
- What is paramount is, instead, that the model provide an out-of-sample prediction that is as accurate as possible.
  - For the in-sample model to be useful out-of-sample, the out-of-sample period (near future) must be like the in-sample data (the historical data) – a condition called *stationarity*...

Stationarity

A time series Y_t is *stationary* if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{s+1}, Y_{s+2}, ..., Y_{s+T})$ does not depend on s; otherwise, Y_t is said to be *nonstationary*. A pair of time series, X_t and Y_t, are said to be jointly *stationary* if the joint distribution of $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, ..., X_{s+T}, Y_{s+T})$ does not depend on s.

Stationarity requires the future to be like the past, at least in a probabilistic sense, i.e., history is relevant!

Stationarity is a key requirement for external validity of a forecast made using time series data.

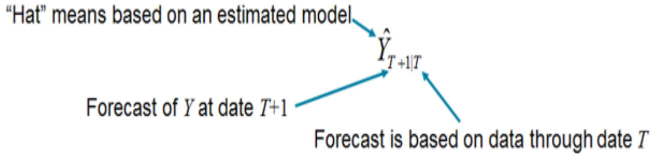
Throughout this lecture, we assume that Y_t is stationary (nonstationary time series will be studied in ECON7350).

Forecasts and Forecast Errors

A *forecast* is a prediction of the future based on time series data. Because the forecasting model is estimated using historical data, a forecast is an out-of-sample prediction of the future.

Forecasts can either be one-step ahead (next month, if you are using monthly data), or multiple steps ahead (4 months from now).

Let the estimation sample be $t = 1, ..., T$. The one-step ahead forecast is:



$$\text{Forecast error} = Y_{T+1} - \hat{Y}_{T+1|T}$$

Forecasts: Terminology and Notation

Predicted values are “in-sample” (the usual definition).

Forecasts are “out-of-sample” - in the future.

Notation:

- $Y_{T+1|T} =$ forecast of Y_{T+1} based on $Y_T, Y_{T-1}, \ldots$ using the population (true unknown) coefficients.
- $\hat{Y}_{T+1|T} =$ forecast of Y_{T+1} based on $Y_T, Y_{T-1}, \ldots$ using the estimated coefficients, which are estimated using data through period T.

A multi-step ahead forecast is for a date more than one period in the future. Specifically, an h-period ahead forecast, made using data through date T, is denoted, $\hat{Y}_{T+h|T}$.

Mean Squared Forecast Error

The *Mean Squared Forecast Error* (MSFE) is a measure of the quality of a forecast.

Specifically, the MSFE is the expected value of the square of the forecast error, when the forecast is made for an observation not used to estimate the forecasting model (i.e., for an observation in the future).

- Using the square of the error penalizes large forecast error much more heavily than small ones.
- A small error might not matter, but a very large error could call into question the entire forecasting exercise.

$$\mathrm{MSFE} = E\left[\left(Y_{T+1} - \hat{Y}_{T+1|T}\right)^2\right]$$

Autoregressions

A natural starting point for a forecasting model is to use past values of Y (that is, $Y_{t-1}, Y_{t-2}, \ldots$) to forecast Y_t.

An *autoregression* (AR) is a regression model in which Y_t is regressed against its own lagged values.

The number of lags used as regressors is called the *order* of the autoregression.

- In a *first order autoregression*, Y_t is regressed against Y_{t-1}.
- In a *pth order autoregression*, Y_t is regressed against $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$.

The AR(1) Model

The population AR(1) model is

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

- β_0 and β_1 do not have causal interpretations.
- If $\beta_1 = 0$, Y_{t-1} is not useful for forecasting Y_t.
- The AR(1) model can be estimated by an OLS regression of Y_t against Y_{t-1} (mechanically, how would you run this regression?)
- Testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ provides a test of the hypothesis that Y_{t-1} is not useful for forecasting Y_t.

Example: AR(1) Model for the Growth Rate of GDP

AR(1) estimated using data from 1962:Q1 – 2017:Q3:

$$\widehat{\mathtt{GDPGR}}_t = \underset{(0.322)}{1.950} + \underset{(0.073)}{0.341} \times \mathtt{GDPGR}_{t-1}, \ \bar{R}^2 = 0.11$$

where $\mathtt{GDPGR}$ is the percentage growth of GDP at an annual rate using the log approximation = $400 \times \ln(\mathtt{GDP}_t/\mathtt{GDP}_{t-1})$.

- Is the lagged growth rate of GDP a useful predictor of the current growth rate of GDP? $|t| = 0.341/0.073 = 4.67 > 1.96$.
- Yes, the lagged growth rate of GDP is a useful of the current growth rate, but the $\bar{R}^2$ is pretty low.
- $\mathtt{GDPGR}_{2017:Q3}$ = 3.11 (units are %). The forecast of $\mathtt{GDPGR}_{2017:Q4}$ is:

$$\widehat{\mathtt{GDPGR}}_{2017:Q4|2017:Q3} = 1.950 + 0.341 \times 3.11 = 3.0$$

The AR(p) Model

The pth order autoregressive model (AR(p)) is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$$

- The AR(p) model uses p lags of Y as regressors.
- The AR(1) model is a special case with $p = 1$.
- The coefficients do not have a causal interpretation.
- To test the hypothesis that Y_{t-2}, ..., Y_{t-p} do not further help forecast Y_t, beyond Y_{t-1}, use an F-test.
- Use t- or F-tests to determine the lag order p.
- Or, better, determine p using an "information criterion" (more on this later...).

Example: AR(2) Model for the Growth Rate of GDP

AR(2) estimated using data from 1962:Q1 – 2017:Q3:

$$\widehat{\texttt{GDPGR}}_t = \underset{(0.37)}{1.60} + \underset{(0.08)}{0.28} \times \texttt{GDPGR}_{t-1} + \underset{(0.08)}{0.18} \times \texttt{GDPGR}_{t-2}, \ \bar{R}^2 = 0.14$$

- To test the hypothesis that Y_{t-2} doe not further help forecast Y_t, beyond Y_{t-1}, use the t-test: $0.18/0.08 = 2.30 > 1.96$.
- The $\bar{R}^2$ increases slightly relative to the AR(1) and the second lag is statistically significant. Still, most of the variation in next-quarter's GDP growth remains unforecasted. We would like to do better!

The Autoregressive Distributed Lag (ADL) Model

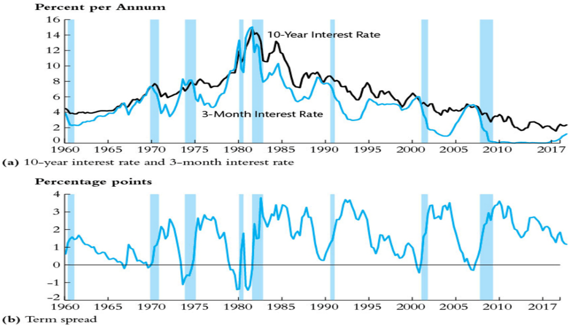
So far we have considered forecasting models using only past values of Y.

It makes sense to add other variables (X) that might be useful predictors of Y, above and beyond the predictive value of lagged values of Y:

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

This is an *autoregressive distributed lag model* with p lags of Y and r lags of X denoted, ADL(p, r).

Example: Interest Rates and the Term Spread



Long-term and short-term interest rates on bonds move together but not one-for-one. The difference between long-term rates and short term rates is called the term spread. The term spread has fallen sharply before U.S. recessions, which are shown as shaded regions in the figures.

Example: Interest Rates and the Term Spread

ADL(2,1) and ADL(2,2) models (1962:Q1 - 2017:Q3):

$$\begin{aligned}\widehat{\texttt{GDPGR}}_t = &\underset{(0.47)}{0.94} + \underset{(0.08)}{0.27} \times \texttt{GDPGR}_{t-1} + \underset{(0.08)}{0.19} \times \texttt{GDPGR}_{t-2} \\ &+ \underset{(0.18)}{0.42} \times \texttt{TSpread}_{t-1}, \ \bar{R}^2 = 0.16\end{aligned}$$

$$\begin{aligned}\widehat{\texttt{GDPGR}}_t = &\underset{(0.46)}{0.94} + \underset{(0.08)}{0.25} \times \texttt{GDPGR}_{t-1} + \underset{(0.08)}{0.18} \times \texttt{GDPGR}_{t-2} \\ &- \underset{(0.42)}{0.13} \times \texttt{TSpread}_{t-1} + \underset{(0.43)}{0.62} \times \texttt{TSpread}_{t-2}, \ \bar{R}^2 = 0.16\end{aligned}$$

The ADL(2,1) model has the first lag of `TSpread` significant, however the two lags in the ADL(2,2) model are not significant.

Estimation of the MSFE and Forecast Intervals

Why do you need a measure of forecast uncertainty?

- To construct forecast intervals.
- To let users of your forecast (including yourself) know what degree of accuracy to expect

Consider the forecast

$$\hat{Y}_{T+1|T} = \hat{\beta}_0 + \hat{\beta}_1 Y_T + \hat{\delta}_1 X_T$$

The forecast error is:

$$Y_{T+1} - \hat{Y}_{T+1|T} = u_{T+1} - [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1) Y_T + (\hat{\delta}_1 - \delta_1) X_T]$$

The Mean Squared Forecast Error (MSFE)

$$
\begin{aligned}
\mathrm{MSFE} &= E\left[\left(Y_{T+1} - \hat{Y}_{T+1|T}\right)^2\right] \\
&= E[u_{T+1}^2] + E\left\{\left[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T + (\hat{\delta}_1 - \delta_1)X_T\right]^2\right\}
\end{aligned}
$$

The first term, $E[u_{T+1}^2]$ is the MSFE of the oracle forecast – that is, the forecast you would make if you knew the true value of the parameters.

- If the sample size is large and the number of predictors is small, then most of the MSFE arises from this term.
- Under stationarity, $E[u_{T+1}] = 0$, so $E[u_{T+1}^2] = Var(u_{T+1})$.
- In general, however, the forecast error might have a nonzero mean.

The second term arises because you need to estimate the β's.

- If the number of predictors is moderate or large relative to the sample size, then this term can be large, in fact it can dominate the first term.

The Root Mean Squared Forecast Error (RMSFE)

$$\text{RMSFE} = \sqrt{E\left[(Y_{T+1} - \hat{Y}_{T+1|T})^2\right]}$$

- The RMSFE is a measure of the spread of the forecast error distribution.
- The RMSFE is like the standard deviation of u_t, except that it explicitly focuses on the forecast error using estimated coefficients, not using the population regression line.
- The RMSFE is a measure of the magnitude of a typical forecasting "mistake".

Three Ways to Estimate the MSFE

$$\begin{aligned}\mathrm{MSFE} &= E\left[(Y_{T+1} - \hat{Y}_{T+1|T})^2\right] \\ &= E[u_{T+1}^2] + E\left\{[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)Y_T + (\hat{\delta}_1 - \delta_1)X_T]^2\right\}\end{aligned}$$

Method 1: Approximate the MSFE by the SER

- If the data are stationary then the forecast error will have mean zero, so $E[u_{T+1}^2] = \sigma_u^2$.
- If the number of predictors is small compared to the sample size, the contribution of estimation error can be ignored.
- Use the approximation $\approx E[u_{T+1}^2] = \sigma_u^2$, so estimate the MSFE by the square of the standard error of the regression (SER). Let SSR denote the (in-sample) sum of squared residuals:

$$\widehat{\mathrm{MSFE}}_{SER} = s_{\hat{u}}^2 = \frac{SSR}{T-p-1}$$

Three Ways to Estimate the MSFE

Method 2: Estimate the MSFE using the final prediction error (FPE)

The Final Prediction Error is an estimate of the MSFE that incorporates both terms in the MSFE formula (squared future u and contribution from estimating the coefficients), under the additional assumption that the error are homoskedastic.

It is shown in SW Appendix 19.7 that this leads to,

$$\widehat{\mathrm{MSFE}}_{FPE} = \left(\frac{T+p+1}{T}\right) s_{\hat{u}}^2 = \left(\frac{T+p+1}{T-p-1}\right)\left(\frac{SSR}{T}\right)$$

Three Ways to Estimate the MSFE

Method 3: Estimate the MSFE by pseudo out-of-sample (POOS) forecasting

- The SER method of estimating the MSFE assumes stationarity and ignores estimation error.
- The FPE method of estimating the MSFE incorporates estimation error, but requires stationarity and homoskedasticity.
- The third method for estimating the MSFE incorporates the estimation error and requires neither stationarity nor assumptions like homoskedasticity.
- This method, pseudo out-of-sample (POOS) forecasting, is based on simulating how the forecasting model would have done, had you been using it in real time.

Three Ways to Estimate the MSFE

Method 3: Estimate the MSFE using by pseudo out-of-sample (POOS) forecasting

1. Choose a date P for the start of your POOS sample, for example $P = 0.1T$ or $0.2T$. Let $s = T - P$.
2. Estimate the forecasting model using the estimation sample, i.e., observations $t = 1, ..., s$.
3. Compute your POOS forecast for date $s + 1$, using the model estimated through s. This is $\hat{Y}_{s+1|s}$.
4. Compute the POOS forecast error, $\tilde{u}_{s+1} = Y_{s+1} - \hat{Y}_{s+1|s}$.
5. Repeat 2-4 for the remaining periods, $s = T - P + 1, ..., T - 1$ and compute

$$\widehat{\mathrm{MSFE}}_{POOS} = \frac{1}{P} \sum_{s=T-P+1}^{T} \tilde{u}_s^2$$

Using the RMSFE to Construct Forecast Intervals

If u_{T+1} is normally distributed, then a 95% forecast interval can be constructed as

$$\hat{Y}_{T+1|T} \pm 1.96 \times \widehat{\mathrm{RMSFE}}$$

Note:

- A 95% forecast interval is not a confidence interval (Y_{T+1} is not a nonrandom coefficient, it is random!)
- This interval is only valid if u_{T+1} is normal - but still might be a pretty good approximation, and it is a commonly used measure of forecast uncertainty.
- Frequently, "67%" forecast intervals are used: $\hat{Y}_{T+1|T} \pm \widehat{\mathrm{RMSFE}}$.

Lag Length Selection Using Information Criteria

How to choose the number of lags p in an AR(p)?

- Omitted variable bias is irrelevant for forecasting!
- You can use sequential "downward" t- or F-tests; but the models chosen tend to be "too large". (why?)
- Another - better - way to determine lag lengths is to use an *information criterion* (IC).
- IC trade off bias (too few lags) vs. variance (too many lags).
- Two IC are the Bayes (BIC) and Akaike (AIC)...

The Bayes Information Criterion (BIC)

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

- First term: always decreasing in p (larger p, better fit).
- Second term: always increasing in p.
  - The variance of the forecast due to estimation error increases with p - so you do not want a forecasting model with too many coefficients - but what is "too many"?
  - This term is a "penalty" for using more parameters - and thus increasing the forecast variance.
- Minimizing BIC(p) trades off bias and variance to determine a "best" value of p for your forecast.

It can be shown that $\hat{p}^{BIC} \xrightarrow{p} p$ (SW, Appendix 15.5).

Another Information Criterion: Akaike Information Criterion (AIC)

$$\text{AIC}(p) = \ln\left(\frac{SSR(p)}{T}\right) + (p+1)\frac{2}{T}$$

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

The penalty term is smaller for AIC than BIC ($2 < \ln T$).

- AIC estimates more lags (larger p) than the BIC.
- This might be desirable if you think longer lags might be important.
- However, the AIC estimator of p isn't consistent - it can overestimate p - the penalty is not big enough.

Example: AR model of GDP Growth, lags 0 - 6

TABLE 15.3: The Bayes Information Criterion (BIC) and the R^2 for Autoregressive Models of U.S. GDP Growth Rates, 1962:Q1–2017:Q3

P	$SSR(p)/T$	$\ln[SSR(p)/T]$	$(p+1)\ln(T)/T$	$BIC(p)$	R^2
0	10.477	2.349	0.024	2.373	0.000
1	9.247	2.224	0.048	2.273	0.117
2	8.954	2.192	0.073	2.265	0.145
3	8.954	2.192	0.097	2.289	0.145
4	8.920	2.188	0.121	2.310	0.149
5	8.788	2.173	0.145	2.319	0.161
6	8.779	2.172	0.170	2.342	0.162

- BIC chooses 2 lags.
- If you used the R^2 to enough digits, you would (always) select the largest possible number of lags

Generalization of BIC to Multivariate (ADL) Models

Let K = the total number of coefficients in the model (intercept, lags of Y, lags of X). The BIC is,

$$\mathrm{BIC}(K) = \ln\left(\frac{SSR(K)}{T}\right) + K\frac{\ln T}{T}$$

- Can compute this over all possible combinations of lags of Y and lags of X (but this is a lot)!
- In practice you might choose lags of Y by BIC, and decide whether or not to include X using a Granger causality test with a fixed number of lags (number depends on the data and application).

Summary: Time Series Forecasting Models

The tools of regression can be used to construct reliable forecasting models - even though the coefficients do not need to be causal for forecasting purposes, and rarely are:

- AR(p) - common "benchmark" models.
- ADL(p, q) - add q lags of X (another predictor).
- Granger causality tests - test whether a variable X and its lags are useful for predicting Y given lags of Y.

New ideas and tools:

- stationarity.
- forecast intervals using the RMSFE.
- pseudo out-of-sample forecasting.
- AIC/BIC for model selection.