Stat 443. Time Series and Forecasting.

Main ideas:

Time series forecasting with explanatory variables.

Cross-correlations between two series.

Example: Canadian monthly GDP (gross domestic product) and unemployment rate (data set at course web site).

Can forecast of unemployment rate be improved using recent past values of both monthly GDP and unemployment rate?

Can forecast of GDP rate be improved using recent past values of both monthly GDP and unemployment rate?

Suppose we have a bivariate time series $(y_{1,t}, y_{2,t})$, $t = 1, \ldots, T$, or (x_t, y_t).
$\mathcal{F}_{t-1}$ is used for information (observed data) up to time $t - 1$.
A model of form

$$\mathsf{E}\left(Y_t | \mathcal{F}_{t-1}, x_t\right) = \alpha + \phi y_{t-1} + \beta x_t$$

is descriptive, and uses the cross-sectional dependence of x_t with y_t. But it cannot be used for forecasting.

For forecasting, one would have to consider models of form:

$$\mathsf{E}\left(Y_t | \mathcal{F}_{t-1}\right) = \alpha + \phi y_{t-1} + \beta x_{t|t-1}$$

$$\text{or } \mathsf{E}\left(Y_t | \mathcal{F}_{t-1}\right) = \alpha + \phi y_{t-1} + \gamma x_{t-1}.$$

The latter is useful if x at a previous time $t - 1$ influences y at time t. The cross-correlation function can be used to assess whether this assumption is reasonable.

See Sections 5.4, 5.6 of textbook H&A.

Example. Consulting project in 1990s.
Ozone forecasting (May to September, Lower Fraser Valley and Greater Vancouver). Training set: 6 years (May to September) of data.
Response variables: maximum hourly ozone concentration at six stations in Lower Fraser Valley.
Regression models: weighted least squares.

Explanatory variables in final regression model (several variables associated with ozone level, based on meteorology):

- previous day's maximum ozone concentration (ppb);

- temperature (Celsius);

- indicator of precipitation (rain);

- indicator of positive pressure difference relative to Vancouver airport;

- Julian day (day since beginning of year, for quadratic trend)

Forecasting means that meteorological variables (temperature and two indicators) from the same day are not available but could be forecast one day in advance, so that high ozone warnings for the next day could be given one day in advance.

Time series modeling for GDP and unemployment rate

In general, consider two macro-economic variables: can one variable be considered a leading indicator for the other variable. Does GDP = gross domestic product lead unemployment rate or does uneployment rate lead GDP?

Label that two variables as (y_1, y_2) with observations of a bivariate time series $(y_{1,t}, y_{2,t})$, $t = 1, \ldots, T$.
Convert to the differenced series

$$(y'_{1,t}, y'_{2,t}) = (y_{1,t} - y_{1,t-1}, \, y_{2,t} - y_{2,t-1}), \quad t = 2, \ldots, T,$$

or (for many positive-valued macroeconomic variables)

$$(y'_{1,t}, y'_{2,t}) = (\log(y_{1,t}) - \log(y_{1,t-1}), \, \log(y_{2,t}) - \log(y_{2,t-1})), \, t = 2, \ldots, T$$

The two series are (strongly) correlated if the changes in one series are (strongly) correlated with changes in the other series. For example, increases in GDP are associated with decreases in unemployment rate.

Visualizations to understand whether one variable leads the other: plots with lagged variables.

$(z_t : t = 1, \ldots, T)$: `lag(z,k)` for k=1 has $(z_2, \ldots, z_T, NA)$, `lag(z,k)` for k=-1 has $(NA, z_1, \ldots, z_{T-1})$. (Add k to subscript).

Compare correlations of
$\{(y'_{1,t}, y'_{2,t-1}), t = 3, \ldots, T\}$ (lag $+1$ for y'_1) and
$\{(y'_{1,t-1}, y'_{2,t}), t = 3, \ldots, T\}$ (lag -1 for y'_1)

Cross-correlation: pages 19 and 23; Shumway and Stoffer (2017). *Time Series Analysis and Its Applications*, 4th ed, Springer.
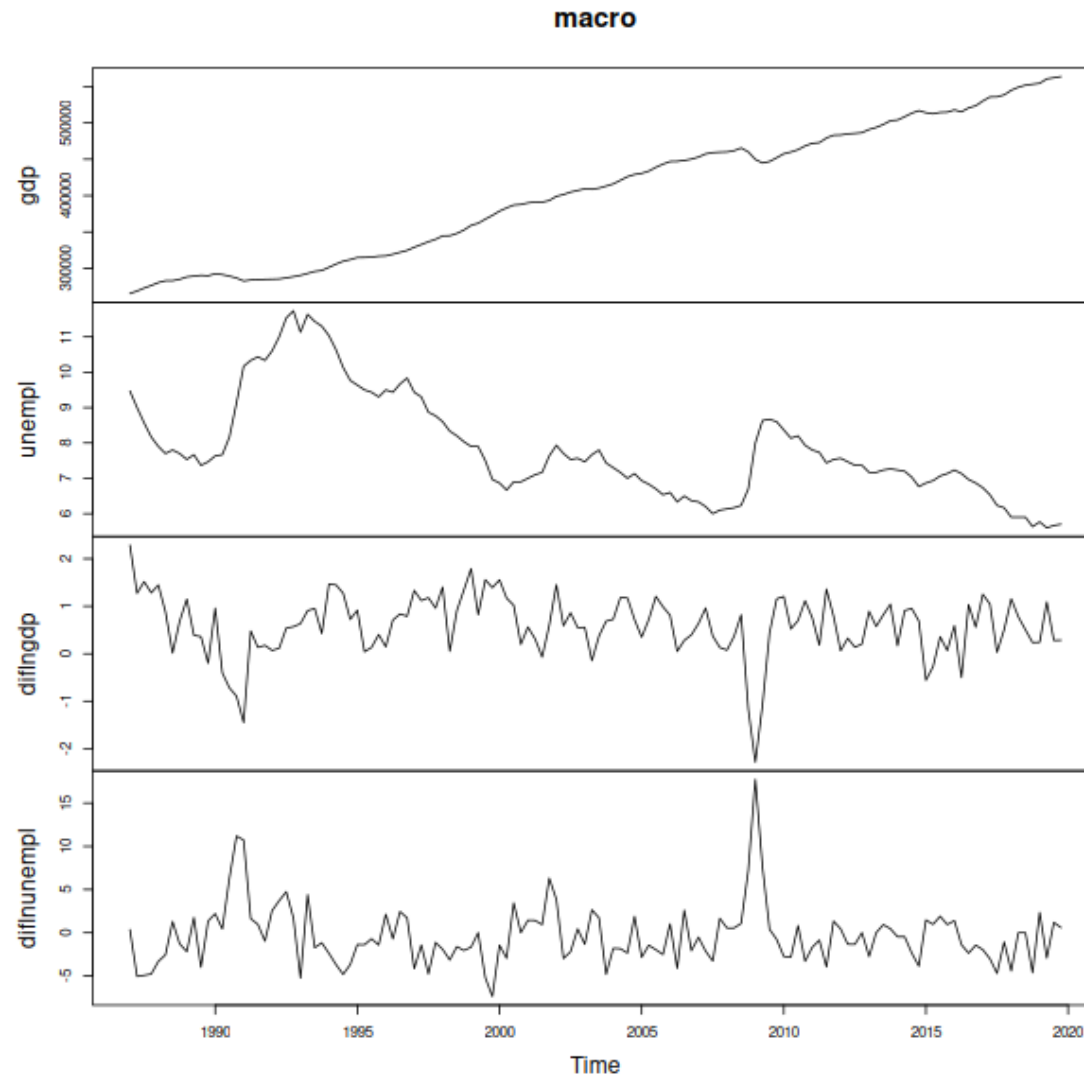
Variables x, y; data $(x_1, y_1), \ldots, (x_T, y_T)$ realizations of $(X_1, Y_1), \ldots, (X_T, Y_T)$).
$\gamma_{xy}(s, t) = \mathrm{Cov}\,(X_s, Y_t)$, $\rho_{xy}(s, t) = \gamma_{x,y}(s, t)/[\gamma_{xx}(s, s)\gamma_{yy}(t, t)]^{1/2}$.
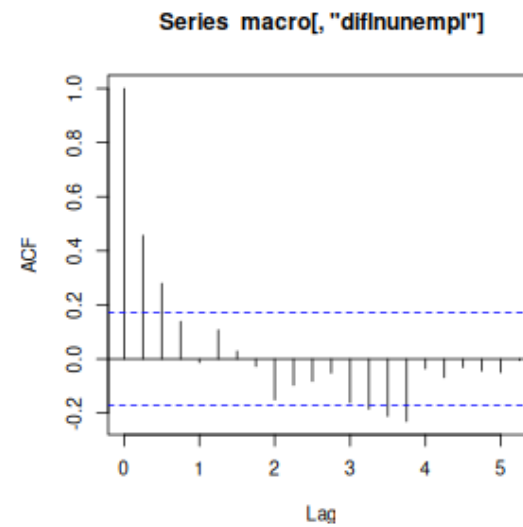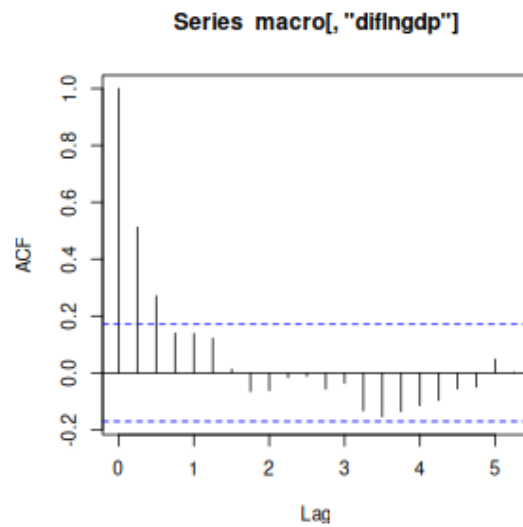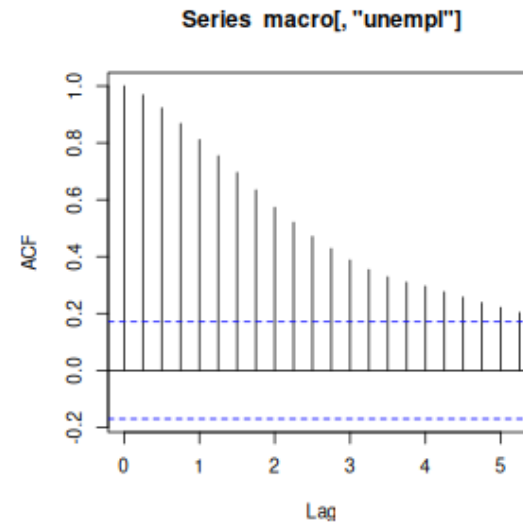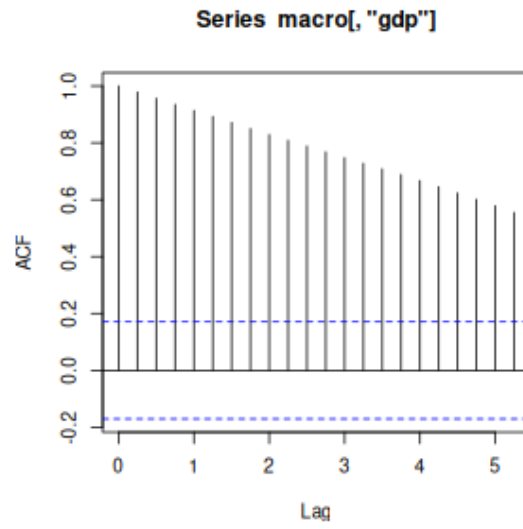If stationary, these depend on s, t only through $t - s$.

cross-autocovariance function: redefine $\gamma_{xy}(s + h, s)$ as $\gamma_{xy}(h)$,
cross-correlation function:

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{[\gamma_x(0)\gamma_y(0)]^{1/2}}.$$
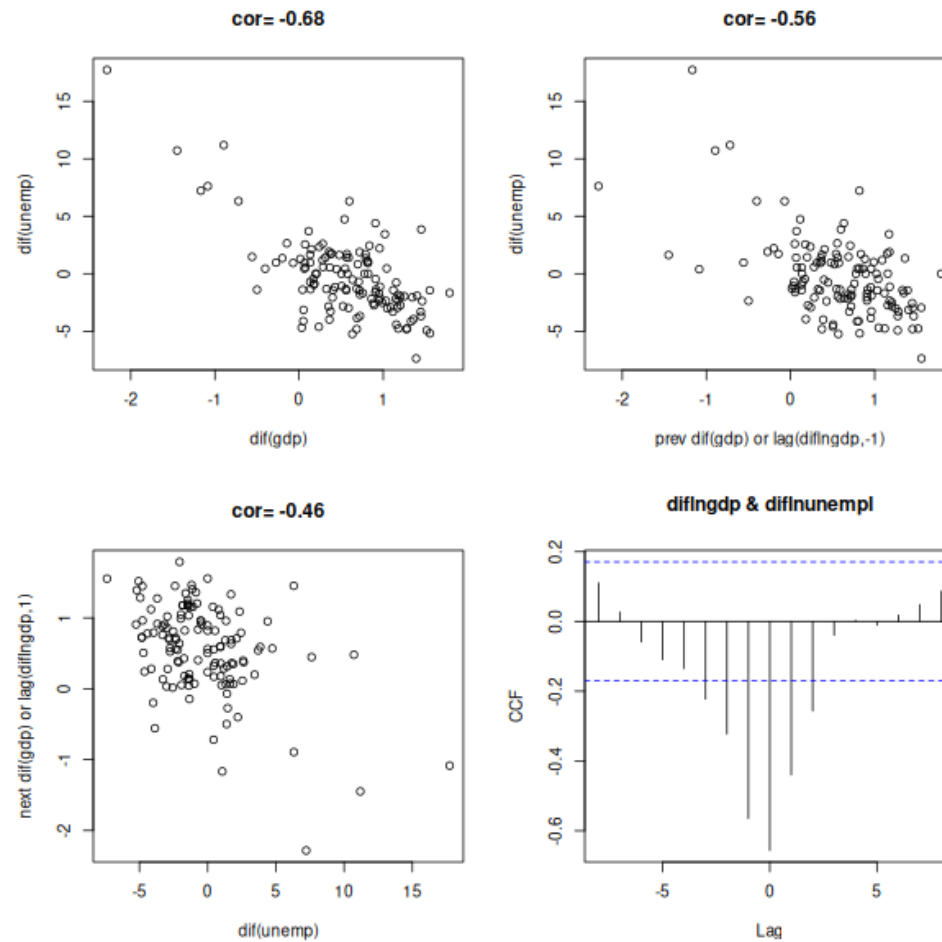
Note that $\gamma_{xy}(h)$ and $\gamma_{xy}(-h)$ are different for $h \neq 0$.

`ccf` in R has the sample cross-correlation function.

macro

Series macro[, "gdp"]

Series macro[, "unempl"]

Series macro[, "diflngdp"]

Series macro[, "diflnunempl"]

4th plot is from ccf(diflngdp,diflnunempl); what would ccf(diflnunempl,diflngdp) look like?

Interpretation for plots: changes in GDP lead changes in unemployment rate with feedback.

Why?

Numerical values of ccf

lag	-4	-3	-2	-1	0	1	2	3	4
ccf	-0.14	-0.22	-0.32	-0.56	-0.66	-0.44	-0.26	-0.04	0.00

Partial case study in Rmd file: compare forecasting of unemployment: previous forecast rules and the new rule that involves regressing diflnunempl on previous(diflnunempl) and previous(diflngdp).

Data series from https://fred.stlouisfed.org/categories/32268; subseries of 1961-Q1 to 2025-Q1, starting from 1987-Q1 to 2019-Q4. 20 years (80 quarters) for training set; 13 years (52 quarters) for holdout set.

Let x = monthly GDP, y = monthly unemployment rate

The wrangle csv file has diflngdp $= x'_t = 100[\log(x_t) - \log(x_{t-1})]$ and $y'_t = 100[\log(y_t) - \log(y_{t-1})]$ (to avoid leading 0.0..., because of small relative monthly changes).

If y'_t (change in unemployment rate) is regressed on x'_{t-1} (previous change in CDP) and y'_{t-1} (previous change in unemployment rate), then the prediction equation leads to:

$$\widehat{y}'_t = 100[\log(\widehat{y}_t) - \log(y_{t-1}]$$

and

$$\widehat{y}_t = \exp\{\widehat{y}'_t/100 + \log(y_{t-1})\} = y_{t-1}\exp\{\widehat{y}'_t/100\}.$$

Similar for y'_t regressed only on x'_{t-1}.

Comparison of forecast rules; see Rmd file at course web site.

rule	Holt	persist	average	AR1	x'_{t-1}	x'_{t-1}, y'_{t-1}
rmse	0.244	0.252	1.723	0.251	0.220	0.218

The simple persist and average rules provide baselines for comparisons.

Summary:

Time series regression modeling/forecasting involves lagged variables, or forecasts of predictor variables.

The cross-correlation function can provide indicator of which variable leads another for two stationary series.