6.1: Introduction to Time Series

Dr. Bean - Stat 5100

1 Why Time Series?

Recall our basic multiple linear regression model:

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \ldots + \beta_{p-1} X_{t,p-1} + \varepsilon_t \qquad (t \text{ index for time})$$
$$\varepsilon_1, \ldots, \varepsilon_n \ \underline{\text{i}}\text{id} N(0, \sigma^2)$$

Previous diagnostics focused on normality and constant variance, but not so much on *independence*.

Violations of independence sometimes detected by **patterns** in residuals over time.

This dependency is often due to auto-correlation ("self-correlation"), which is when the residuals are correlated *with each other*.

What are some examples where you would expect the residuals of a linear model to be auto-correlated over time?

1.1 Autocorrelation, whats the big deal?

- If a random variable is autocorrelated over time, then observations closer in time will tend to be more similar than observations far away in time.
- Thus, repeated samples of the variable *in* time will have **less** variability within the sample than the variability *across* time.
- This means we will **underestimate** the true variance of the random variable, which in OLS causes
  1. The estimates regression coefficients are unbiased, but no longer "best" (i.e. minimum variance)
  2. MSE will underestimate the true residual variance
  3. OLS may also underestimate $s\{b_k\}$, which makes the t-tests unreliable (i.e. destroys inference)

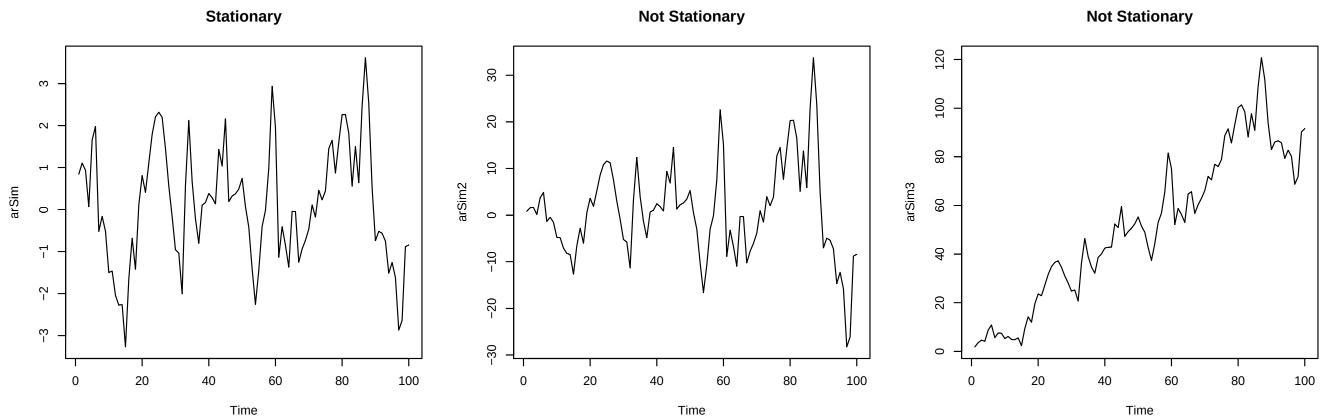

Figure 1: Examples of stationary and non-stationary time series.

2 Time Series Modeling

- autocorrelation means our data contain *structure* over time.
- Accounting for this structure should improve our ability to predict.
- One approach: Box-Jenkins (ARIMA) time series modeling:
  1. Make data stationary
  2. Test for independence
  3. Use sample autocorrelation and sample partial autocorrelation plots to identify potential dependence structures
  4. Fit dependence structures and asses model adequacy
  5. Using adequate model
     - Forecase response variable (w/ confidence interval)
     - Test model terms (incl. predictor variables)

1. **Make data stationary:**

   - First Order (constant mean):
   $$E[\epsilon_t] = \mu_t \equiv \mu \text{ for all } t$$
   - Second Order (constant variance):
   $$Var[\epsilon_t] = \sigma_t^2 \equiv \sigma^2 \text{ for all } t$$
   - This means that if both conditions are satisfied, the time series will "look" the same no matter what time window (with appropriate scale) that we look at.
     - Graphical check: plot residuals e_t vs t (see Figure 1):
     - SAC (sample autocorrelation; ACF) plot - coming up, a useful diagnostic for stationarity
   - Remedial Measures for Non-Stationarity
     - Non constant variance $\rightarrow$ transform Y_t

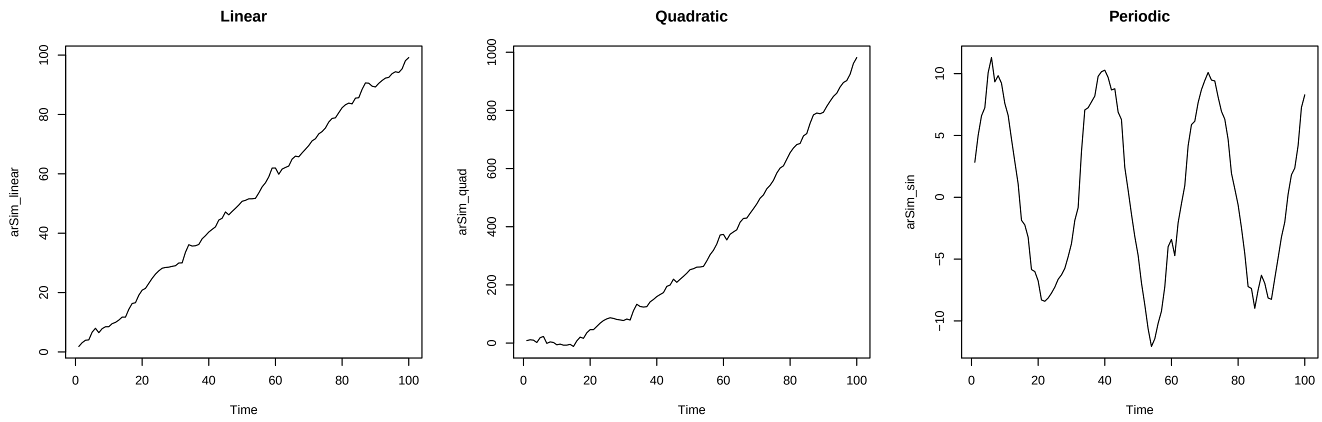

Figure 2: Examples of different trends that may occur in a time series.

- Non-constant mean: "de-trend" the data using a predictive model where time is the explanatory variable.
  * Use a scatter-plot of time vs residuals to determine an appropriate model (see Figure 2).
- "Differencing" for stubborn trends:
  * First differences: $Z_t = Y_t - Y_{t-1}, \quad (t = 2, \ldots, n)$
  * Second differences: $W_t = Z_t - Z_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2} \quad (t = 3, \ldots n)$
- HOWEVER, differencing will make periodic cycles unrecoverable, which can hurt our ability to make forecasts.
- For this reason, differencing is a remedial measure of last resort.

2. **Test for independence**

   There is a difference in a series being a function of time (plus random noise) versus a series that is *correlated* in time.

   Question: Failing to remove time-dependent trends in our data ruins our ability to check for time-dependent correlations, why is this?

   AFTER removing trends, determine if the data are just "white noise" (no dependence structure)
   H_0 : Data are just white noise

   in SAS: χ^2 test for lags 1 through k, (where k is selected by the user).

3. **Identify tentative dependence structures**

   - Notation: Z_t is the stationary time series after "transforming" (including estimating out time trends and other covariates) the original time series $Y_1, \ldots, Y_n$

   - Sample autocorrelation function (ACF or SACF)

$$r_m = \text{linear association (correlation) between time series observations separated by a lag of } m \text{ time units}$$

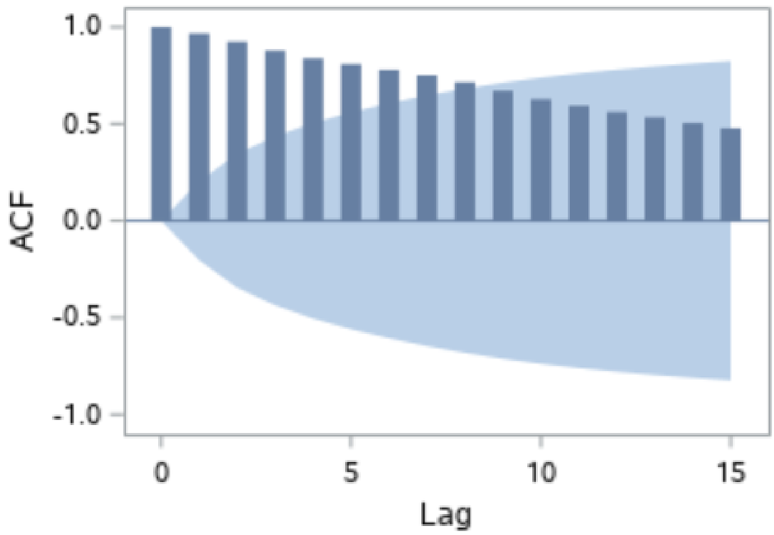

Figure 3: Sample ACF plot for a non-stationary time series.

- PLOT 1: sample autocorrelation plot (or SAC / ACF): check for stationarity and identify tentative dependence structure
  - bar-plot r_m vs. m for various lags m
  - lines often added to represent 2 SE's (rough significance threshold)
  - SAC / ACF terminology:
    * "spike" : r_m is "significant"
    * "cuts off" : no "significant" spikes after r_m
    * "dies down" : decreases in "steady fashion"
  - If Z_t stationary, SAC either cuts off fairly quickly or dies down fairly quickly (sometimes in "damped exponential" fashion)
  - If SAC dies down extremely slowly, Z_t nonstationary (see Figure 3)

- Sample partial autocorrelation function (PACF or SPACF)

$$r_{m,m} \;=\; \text{autocorrelation of time series observations separated by a lag of } m \text{ with the effects of the intervening observations eliminated}$$

- PLOT 2: sample partial autocorrelation plot (or SPAC / PACF)
  - bar-plot $r_{m,m}$ vs. m for various lags m
  - lines often added to represent 2 SE's (rough significance threshold)

- Main dependence structures

  (a) AR(p) dependence structure: autoregressive process of order p:
  - current time series value depends on past values; common representation for AR(p):

$$Z_t \;=\; \delta + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t$$

* ϕ_i are unknown parameters; random shock a_t iid $N(0, \sigma^2)$

- identify using SPAC: first p terms of SPAC will be non-zero, then drop to zero (sketch)

(b) MA(q) dependence structure: moving average process of order q:

- current time series value depends on previous random shocks
- model:

$$Z_t = \delta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Z_t : stationary "transformed" time series θ_i : unknown parameters
a_t : random shocks δ : unknown parameter

- identify using SAC: first q terms of SAC will be non-zero, then drop to zero (sketch)

Common Dependence Structures for Stationary Time Series

	SAC	SPAC
MA(1)	cuts off after lag 1	dies down, dominated by damped exponential decay
MA(2)	cuts off after lag 2	dies down, in mixture of damped exp. decay & sine waves
AR(1)	dies down in damped exponential decay	cuts off after lag 1
AR(2)	dies down, in mixture of damped exp. decay & sine waves	cuts off after lag 2
ARMA(1,1)	dies down in damped exp. decay	dies down in damped exp. decay

ARIMA(p,d,q) dependence structure: Autoregressive **Integrated** Moving Average Model

- a very flexible family of models $\Rightarrow$ useful prediction
- recall first difference: $Z_t = Y_t - Y_{t-1}$, $t = 2, \ldots, n$
  and second difference: $W_t = Z_t - Z_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2}, t = 3, \ldots, n$
- after differencing, AR and MA dependence structures may exist: ARIMA(p, **d** ,q)
  - p : AR(p) – value at time t depends on previous p values)

  - d : # of differences (need to take d^{th} difference to make stationary)
  - q : MA(q) – value at time t depends on previous q random shocks)
- use SAC and SPAC to select p and q – but how to select d?
  - usually look at plots of time series
  - choose lowest d to make stationary (also SAC)
- sometimes see backshift notation: $BY_t = Y_{t-1}$
  - $d = 1$: $Z_t = Y_t - Y_{t-1} = Y_t - BY_t = (1 - B)Y_t$
  - general d: $Z_t = (1 - B)^d Y_t$

- "Fit model" → estimates & standard errors for β_j's, ϕ_l's, & θ_l's

- Several approaches exist to estimate ϕ_l's, θ_l's, and β_j's, and deal with initial lag; we'll use ULS (unconditional least squares) for MA(q) & AR(p)

- ARIMA(p,d,q) model rewritten, with $t = 1, \ldots, n$:

$$\begin{aligned} Y_t &= g_1(Y_1, \ldots, Y_{t-1}) + g_2(X_{t,1}, \ldots, X_{t,k-1}) + g_3(a_1, \ldots, a_t) \\ &\text{where} \\ g_1 &= \text{linear combination (LC) of previous observations} \\ g_2 &= \text{LC of predictors at time } t\text{, in terms of parameters } \beta_j \\ g_3 &= \text{function of random shocks in terms of parameters } \phi_l \ \& \ \theta_l \end{aligned}$$

4. **Fit dependence structures and assess model adequacy**

- General SAS code for ARIMA(p,d,q), Y in terms of $X_1, \ldots, X_{k-1}$:

```
proc arima data = a1;
 identify var = Y (d )  crosscorr = (X1 ... Xk-1) ;
 estimate p = p  q = q  input = (X1 ... Xk-1)  method = uls  plot;
 forecast lead = L alpha = a noprint out = fout;
run;
```

option	description
d, p, q	differencing, AR, & MA settings (as before)
`plot`	adds RSAC & RSPAC plots
L	# times after last observed to forecast
a	set confidence limit; a = .10 ⇒ 90% conf. limits
`noprint`	optional, suppresses output
`out = fout`	optional, sends forecast data to `fout` data set

- Useful diagnostics for "goodness of fit":
  - Numerical

* Standard Error – measure of "overall fit"; in SAS: `Std Error Estimate`

$$S = \sqrt{\frac{\sum_1^n (Y_t - \hat{Y}_t)^2}{n - n_p}}, \qquad n_p = \#\text{ parameters in model}$$

* Ljung-Box statistic Q^* (& p-value);
  in SAS: lag 6 χ^2 for `Autocorrelation Check of Residuals`
  · basic idea: look at "local" dependence among residuals in first few sample autocorrelations
  · under H_0: "model is adequate", $Q^* \sim \chi^2_{df}$

– Graphical (PLOTS 3 and 4) – focus on residuals
  * Residual sample autocorrelation plot (RSAC)
  * Residual sample partial autocorrelation plot (RSPAC)

Question: How will we know from these plots if we "succeeded"?

5. **Using adequate model:**
   (a) forecast response (*** w/ conf. interval ***) – careful far beyond data
   (b) test model terms (incl. predictor variables, but also AR & MA parameters)