Time Series Analysis: Process Outline/Summary

SHORT VERSION

0. Load and format data: correct missing values, dummy variables, aggregation.
1. Plot, examine, and correct the as-is data for trends to arrive at a stationary error process, Xt:
    a. Determine deterministic trends and fix using model approaches or data transformations.
    b. Determine stochastic trends and fix using differencing and unit-root hypothesis tests.
2. Analyze standardized residuals to verify 4 key assumptions made on Xt:
    a. Conduct independence tests to check independence of Xt.
    b. Check for zero-mean and homoscedasticity via residual plots.
    c. Check for normality via QQ plots, histograms, and Shapiro-Wilks test.
3. Determine order of appropriate ARIMA(p,d,q) model:
    a. Check ACF, PACF, and EACF plots to determine candidate pairs of (p,q).
    b. Estimate the candidate models using Maximum Likelihood Estimation (MLE).
    c. Choose the most reasonable model with the smallest value of relevant Information Criteria.
4. Conduct parameter estimation for chosen ARIMA(p,d,q) model to determine AR(p) and MA(q) coefficients.
5. Conduct model diagnostics on chosen model to determine if choice of (p,d,q) truly works for the observed data:
    a. Conduct residual analysis on the estimated error process to check for independence, zero-mean, homoscedasticity, and normality.
    b. If chosen ARIMA(p,d,q) was truly a good fit, the additional AR(p) and MA(q) parameters will not be significant and will result in a model with redundant parameters, causing the estimates of the ARIMA(p,d,q) part to become invalid.

EXTENDED VERSION

0. Load and format data: correct missing values, dummy variables, aggregation.

1. Plot, examine, and correct for trends to arrive at the (hopefully stationary? stationarity needed for ARMA modeling) error/stochastic process, X_t:

   - Deterministic Trends (seasonality, linearity)
     - Fix using...
       * Model Approaches: linear trend models, seasonal means models, cosine models
       * Data Transformations: log, percentage change ($\equiv \iff$ stationary)
     - Remove trend using linear trend model (parametric):

$$Y_t = \beta_0 + \beta_1 t + X_t \text{ where } \hat{X}_t = Y_t - \hat{\beta}_0 + \hat{\beta}_1 t$$

   - Stochastic Trends (non-stationarity)
     - Fix using...
       * Differencing. Determine order of integration, d, by performing repeated unit root tests until d is clear. Take d differences.
     - Unit-Root Hypothesis Tests:
       * Augmented Dickey Fuller (ADF), H_0: "Y_t non-stationary."
       * Phillips–Perron (PP), H_0: "Y_t non-stationary."
       * Kwiatkowski–Phillips–Schmidt–Shin (KPSS), H_0: "Y_t stationary."

2. Analyze the *standardized residuals* to verify the 4 key conditions/assumptions made on X_t:

   - Independence* Tests: Needed for regression coefficients to be meaningful. Check independence first, because if X_t independent then $X_t \overset{iid}{\sim}$ and the other 3 assumption checks aren't needed.):
     - Runs (NON-parametric), H_0: "X_t are independent."
     - Ljung-Box / Portmanteau, H_0: "X_t are independent".
     - ACF (for MA(q); indirect and direct effects)
       * Population ACF: Use if zero-mean stationarity is *known/true.* If MA(q), cuts off at lag q. Bounds are $\pm\frac{1.96}{\sqrt{n}}$
       * Sample ACF: Use if stationarity *either unknown or false.*
     - PACF (for AR(p); direct effects)
       * Population PACF: Use if zero-mean stationarity is *known/true.* If AR(p), cuts off at lag p. Bounds are $\pm\frac{1.96}{\sqrt{n}}$
       * Sample PACF: Use if stationarity *either unknown or false.*
     - If independence fails: use the HAC estimator as it relaxes the need for the independence assumption.
   - Zero-Mean Tests: Check via residual plot (or goodness-of-fit tests, irrelevant to this course). Should have *no* observable pattern. *Must have a zero-mean for a linear trend model to be appropriate.*
   - Homoscedasticity Tests: check via residual plot. Should have *no* observable pattern.
   - Normality Tests:
     - QQ Plot (tail behavior), H_0: "X_t is normally distributed." Points should be on the diagonal/quantile line, no light- or heavy-tailed behavior.
     - Histogram (skew), H_0: "X_t is normally distributed." Should closely mirror an ideal ~Normal distribution without skew.
     - Shapiro-Wilks, H_0: "X_t is normally distributed."

3. Once stationary, you can choose to determine the order of an appropriate ARIMA(p, d, q) model.

1. Check ACF, PACF, and EACF plots to determine candidate pairs of (p, q).
2. Estimate the candidate models using Maximum Likelihood Estimation (MLE).
3. Assess the candidate models relative to each other. Choose *the most reasonable model with the smallest value* of the relevant Information Criteria / "metric":
   - Forecasting or Predicting future values of Y_t: AIC and/or AICc
   - Estimating or "Explaining" aspects of the true model of Y_t: BIC

4. Conduct parameter estimation for your chosen ARIMA(p, d, q) model to determine the AR(p) coefficients $(\phi_1, \cdots, \phi_p)$, the MA(q) coefficients $(\theta_1, \cdots, \theta_q)$, and the σ_e^2. This will give us an estimate of the error e_t:
$$\hat{e_t} = Y_t - \hat{\theta}_0 - \hat{\phi}_1 Y_{t-1} - \hat{\phi}_2 Y_{t-2} - \cdots - \hat{\phi}_p Y_{t-p}\,; \text{ for } t = p+1, \ldots, n$$
Only the LSE and MSE methods are recommended:

   - Least Squares Estimation (LSE [CSS]), for $q \neq 0$and for *large* n.
     - *If* Y_t is AR(p) then LSE $\approx$ MoM. Advantage of the LSE (CSS) estimator is that it is well-defined for all MA(q) and ARMA(p, q) models even when $q > 0$.
     - *If* sample size n is "large enough" then LSE $\approx$ MSE.
   - Maximum Likelihood Estimation (MSE), for $q \neq 0$ and for *small* n.
     - Gives more accurate estimates with nice asymptotic properties, but requires numerical approximations and therefore has slower computation than LSE.
     - Performs pretty well when there are a few departures from the ~Normality assumption.
     - *If* sample size n is "large enough" then MSE $\approx$ LSE.
   - Method of Moments (MoM) / Yule-Walker Estimation, only appropriate for ARMA(p, q) model where $q \leq 0$ (!!!).
     - If there is moving-average type dependence in the data, then...
       - [1] MoM may not have a single solution (could have no solutions or multiple solutions).
       - [2] MoM may have unnecessarily high variance.

5. Conduct model diagnostics on the chosen model to determine if your choice of (p, d, q) truly works for the observed data.

   1. Once we have an estimate of the error e_t from our parameter estimation in step 4, we can define our prediction error as $\hat{e_t} = Y_t - \hat{Y_t}$. Then, we *conduct residual analysis on the estimated error process* ("residuals"), $\{\hat{e_t}\}$:
      - For an ARMA(p, q) model, $\{\hat{e_t}\}$ *should behave like an iid process* **if**:
        - [1] our specified ARMA(p, q) is "correct."
        - [2] our specified coefficients are "correct."
      - Independence Tests:
        - Runs test.
        - ACF plot *should* decay to zero after lag 0.
        - ACF *Tests* for some time series $\{u_t\}$. Use the **Sample ACF of residuals** to test H_0: "u_t are uncorrelated." To implement the following in R, use `stats::Box.test()`.
          - Box-Pierce / Portmanteau Test, has that $\sqrt{n} \cdot r_k \overset{D}{\sim} \mathrm{N}(0, 1)$. Not ideal because Box-Pierce/Portmanteau relies on $\sim \chi^2$, which gives fairly poor approximations even for large n.
          - Ljung-Box Test is more "accurate." Holds that $Q_* \overset{D}{\sim} \chi^2_{K-p-q}$; where $K := $ maximum lag, r_k is the sample ACFs of $\hat{e_t}$. We would reject H_0 if $Q_* > $ the *5% upper percentile* of χ^2_{K-p-q}.
          - *Important:* use `stats::tsdiag()` to automatically generate diagnostic plots for an ARIMA(p, d, q) model—gives: standardized residual plot, sample ACF plot for residuals, p-values for the Ljung-Box test *for different K's—HOWEVER*, the last plot for

p-values of the Ljung-Box test is not correct because it incorrectly uses df $= K$ as to the correct df $= K - p - q$. To fix this, use `TSA::tsdiag.Arima()` so that R will recognize that you want the `tsdiag` of an ARIMA(p, d, q) fit.

- Zero-mean and Homoscedasticity Tests: Check via residual plot. Should have *no* observable pattern, scattered around 0.
  - If we're using *standardized residuals* ($\hat{e_t^*} = \frac{\hat{e_t}}{\hat{\sigma_e^2}}$) then by the normality assumption we'd expect to see the majority of the residuals $|\hat{e_t^*}| < 3\,(\approx)$.
- Normality Tests (*Departing from the normality assumption is not as serious as the other assumptions, especially if we have a "large" sample size, n):
  - QQ plot close to ideal ~Normal
  - Histogram close to ideal ~Normal
  - Shapiro-Wilks rejected (so errors are ~Normal).

2. If you chose your ARIMA(p, d, q) based on parameter estimation from step 4, include and analyze the two following *additional* models. Do this to use overfitting as a tool for reassuring yourself that your chosen model is *truly* best for your data:
   - [1] ARIMA$(p + 1, d, q)$: Add one more lag to the AR(p) component.
   - [2] ARIMA$(p, d, q + 1)$: Add one more lag to the MA(q)) component.
3. *If* your chosen ARIMA(p, d, q) was truly a good fit, then the additional AR(p) and MA(q) parameters...
   - ...will NOT be significant.
   - ...will result in a model with redundant parameters. This will cause the estimates of the ARIMA(p, d, q) part to become invalid.