2/16/2010 36-402/608 ADA-II H. Seltman

Handout #10: The Nature of Serial Correlation

1. Violation of the "independent errors" assumption comes in two basic forms: clustering and serial correlation

2. Important fact: The linear model is unbiased in the presence of serial correlation, but the standard errors (and therefore confidence intervals and p-values), are wrong.

3. Definition of a "stationary time series": unchanging mean and variance/covariance structure across time. Remove trends and mean before analysis.

4. Definition of autocorrelation of lag "m": covariance of values at times t and $t + m$ divided by variance of values (covariance at t and t).

5. Uncorrelated errors are called "white noise".

6. Standard set of serial correlation models for stationary time series (after subtracting the mean): ARMA

   (a) Autoregressive order p or AR(p) models:

   $$Y_t = \sum_{i=1}^{p} \alpha_i Y_{t-i} + \epsilon_i$$

   A large error (also called an "innovation" in time series analysis) has a long effect in the future. Autocorrelation drops off exponentially. Partial autocorrelation function graph has p peaks.

   (b) Moving average order q or MA(q) models:

   $$Y_t = \sum_{i=1}^{q} \beta_i \epsilon_{t-i} + \epsilon_i$$

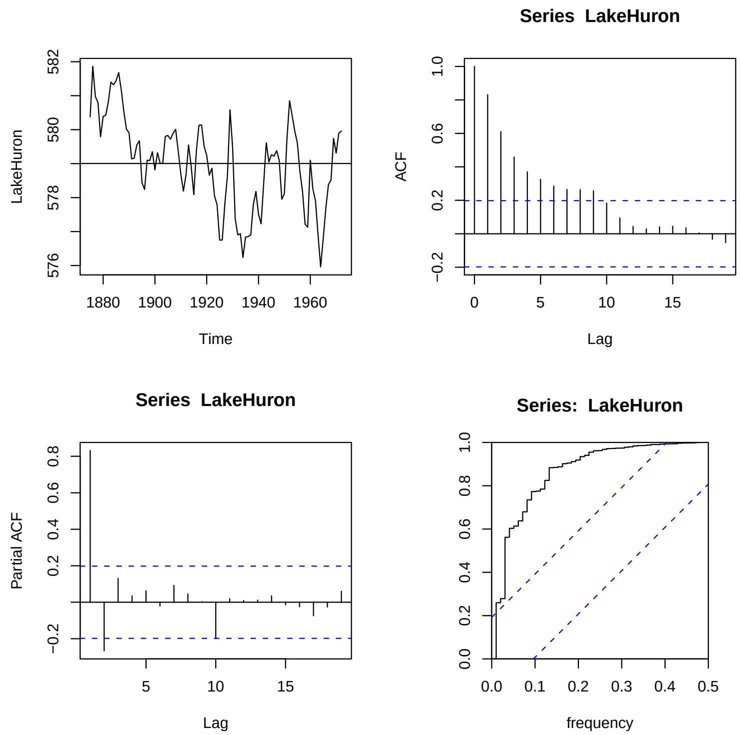
   A large error has a short effect in the future. Autocorrelation drops quickly. Autocorrelation plot shows $1 + q$ peaks.

   (c) ARMA(p,q) models combine both. ARIMA (Integrated) models handle non-stationarity.

   (d) ARISMA models also handle "seasonality".

   (e) Alternative "spectral" models work in the frequency rather than the time domain by using Fourier analysis.

7. Example: "LakeHuron" holds annual measurements of the level, in feet, of Lake Huron 1875-1972.



8. Plots

   (a) A time trace shows “runs” above and below the mean. This suggests falsely low standard errors with positive serial correlation, and great unreliability with very short time series. Non-symmetry of peak shapes suggests the need for a data transformation.

   (b) ACF: autocorrelation function, shows autocorrelation at many time lags; first peak at lag 0, is always 1.0.

   (c) PACF: partial autocorrelation function, is specially designed to identify AR models, showing p large peaks.

   (d) Cumulative periodogram is based on spectral methods, but easily demonstrates the presence of serial correlation.

9. Time series functions

   (a) ts() makes a time series with associated time information.

   (b) ar(data, aic=T, order.max=p) fits AR models up to AR(p) and finds the "best" as the one with lowest AIC.

```
> ar(LakeHuron)
Coefficients:
      1        2
 1.0538  -0.2668
Order selected 2  sigma^2 estimated as  0.5075
> ar(LakeHuron, order.max=4)$aic
          0           1           2           3           4
118.6683709   5.2338642   0.0000000   0.3100411   2.1963067
```

   (c) arima(data, order=c(p,0,q)) fits an ARMA(p,q) model and returns the estimated coefficients (alpha values), sigma squared, and residuals (which can be checked to assure that they are white noise).

```
> arima(LakeHuron, order=c(2,0,0))
Coefficients:
         ar1      ar2  intercept
      1.0436  -0.2495   579.0473
s.e.  0.0983   0.1008     0.3319
sigma^2 estimated as 0.4788:  log likelihood = -103.63,  aic = 215.27
```

   (d) durbin.watson(lm(y x,data=dtf)) in package "car" is a nice test if a regression can be trusted to not have significant serial correlation.

10. Breakout and Discussion