Advanced Methods in Biostatistics II

Lecture 10

November 28, 2017

Linear models with autocorrelated errors

- Consider the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Let us assume that the observations are measured at equally spaced time points and the error terms from adjacent time points are correlated (i.e., autocorrelated).
- This violates the standard assumption of independent errors made in the linear model.

Time series analysis

- To properly understand autocorrelated errors we need some background regarding time series analysis.
- If a random variable X is indexed in time, the observations $\{X_t, t \in T\}$ is called a time series.
- A time series $X_t, t \in T$ can be regarded as a realization of a stochastic process.
- We are in particular interested in discrete equally spaced time series.

Second-order properties

1. The mean function of X_t: $\mu_X(t) = E(X_t)$.
2. The variance function of X_t: $\sigma_X^2(t) = E(X_t - \mu_X(t))$.
3. The autocovariance function of X_t:

$$\gamma_X(r, s) = \mathrm{cov}(X_r, X_s) = E((X_r - E(X_r))(X_s - E(X_s)))$$

for $s, t \in T$.

4. The autocorrelation function of X_t:

$$\rho_X(r, s) = \frac{\gamma_X(r, s)}{\sqrt{\gamma_X(r, r)\gamma_X(s, s)}}.$$

Weak stationarity

A time series X_t is weakly stationary if

1. $E|X_t|^2 < \infty$ for all t.
2. The mean function $\mu_X(t)$ does not depend on t.
3. The covariance function

$$\gamma_X(t, t+h)$$

is independent of t for all h.

Weak stationarity

- If X_t is weakly stationary then the autocovariance function (ACVF) at lag h can be written:

$$\gamma(h) = \mathrm{cov}(X_{t+h}, X_t)$$

- When $h = 0$, we have that $\gamma_X(0) = \mathrm{var}(X_t)$.
- Hence, the autocorrelation function (ACF) of X_t at lag h can be written:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)}.$$

Partial autocorrelation

- In addition to the correlation between X_{t+h} and X_t, we may also want to investigate their mutual dependence after removing the effects of the intervening variables.
- That is we seek to compute the conditional correlation:

$$\phi(h) = \text{corr}(X_t, X_{t+h} | X_{t+1}, \ldots, X_{t+h-1}).$$

- Usually referred to as the partial autocorrelation function (PACF) at lag h.

White noise

- A sequence of uncorrelated random variables Z_t, each with mean 0 and variance σ^2, is called white noise, written $Z_t \sim WN(0, \sigma^2)$.

- A white noise process Z_t has the following properties:

$$E(Z_t) = 0 \quad \forall t,$$

$$Var(Z_t) = \sigma^2 \quad \forall t$$

and

$$\rho(h) = \begin{cases} 1, & \text{if } h = 0 \\ 0, & \text{if } h \neq 0 \end{cases}$$

AR(p) process

- A time series X_t is an autoregressive process of order p, written AR(p):

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$$

where $Z_t \sim WN(0, \sigma^2)$ and $\phi_1, \phi_2, \ldots \phi_p$ are constants.

MA(q) process

- A time series is a moving-average process of order q, written MA(q), if

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}$$

where $Z_t \sim WN(0, \sigma^2)$ and $\theta_1, \theta_2, \ldots \theta_q$ are constants.

ARMA(p,q) model

- A time series is an autoregressive moving-average process, written ARMA(p,q), if

$$
\begin{aligned}
X_t &= \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} \\
&\quad + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q}
\end{aligned}
$$

where $Z_t \sim WN(0, \sigma^2)$ and $\phi_1, \phi_2, \ldots \phi_p, \theta_1, \theta_2, \ldots \theta_q$ are constants.

Characteristics of stationary processes

Process	ACF	PACF
AR(p)	Tails off as exponential decay or damped sine wave	Cuts off after lag p
MA(p)	Cuts off after lag q	Tails off as exponential decay or damped sine wave
ARMA(p,q)	Tails off after lag (q-p)	Tails off after lag (p-q)

Analyzing time series

- Given a set of observations from a stationary time series, the goal of time series analysis is to find an appropriate model to represent the observed data.
- Important issues involve: (i) model selection; (ii) order selection; and (iii) estimation of the model parameters.

Model and order selection

- Candidate models can be identified by studying the ACF and the PACF.
- Model and order selection can be performed using information criteria that assess model fit.
- Here a range of potential models are estimated and a criteria such as AIC or BIC is used to choose the most appropriate.

Analyzing time series

- The parameters of an AR(p) model can be estimated using the Yule-Walker estimates (i.e., method of moments), or alternatively maximum likelihood or restricted maximum likelihood methods.
- The parameters of an ARMA(p,q) model can be estimated using maximum likelihood or restricted maximum likelihood methods.

Method of moments - illustration

- Let us illustrate the method of moments by assuming an AR(2) process:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t$$

- We seek to estimate the parameters ϕ_1, ϕ_2, σ^2.

Method of moments - illustration

- Begin by multiplying the process by X_{t-k} for $k = 1, 2$ and take the expectation.
- This gives the following set of equations:

$$\begin{aligned} E(X_{t-1}X_t) &= \phi_1 E(X_{t-1}X_{t-1}) + \phi_2 E(X_{t-1}X_{t-2}) + E(X_{t-1}Z_t) \\ E(X_{t-2}X_t) &= \phi_1 E(X_{t-2}X_{t-1}) + \phi_2 E(X_{t-2}X_{t-2}) + E(X_{t-2}Z_t). \end{aligned}$$

- Note that $\gamma(k) = E(X_{t-k}X_t)$ and $E(X_{t-k}Z_t) = 0$ for $k \geq 1$.

Method of moments - illustration

- Now divide both equations by $\gamma(0)$.
- This gives the following set of equations:

$$\begin{aligned}\rho(1) &= \phi_1 + \phi_2\rho(1) \\ \rho(2) &= \phi_1\rho(1) + \phi_2\end{aligned}$$

- These are the Yule-Walker estimates.

Sample autocovariance

- From the observations $\{X_1, X_2, \ldots X_n\}$ of a stationary time series X_t we often seek to estimate the autocovariance function $\gamma(.)$.

- The sample autocovariance function is defined by

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{j=1}^{n-h} (x_{j+h} - \bar{x})(x_j - \bar{x})$$

for $0 \leq h \leq n$.

Method of moments - illustration

- Next, compute the sample ACF to obtain $\hat{\rho}(1)$ and $\hat{\rho}(2)$.
- Now equate the sample and population moments and solve these equations.
- This gives the following estimates:

$$\begin{aligned}\hat{\phi}_1 &= \frac{\hat{\rho}(1)(1-\hat{\rho}(2))}{1-\hat{\rho}(1)^2} \\ \hat{\phi}_2 &= \frac{\hat{\rho}(2)-\hat{\rho}(1)^2}{1-\hat{\rho}(1)^2}.\end{aligned}$$

Method of moments - illustration

- To estimate σ^2, multiply the process by X_t and take the expectation.
- This gives the following:

$$\gamma(0) = \phi_1\gamma(1) + \phi_2\gamma(2) + \sigma^2$$

- Solving for σ^2 we obtain the following estimate:

$$\hat{\sigma}^2 = \hat{\gamma}(0)(1 - \hat{\phi}_1\hat{\rho}(1) - \hat{\phi}_2\hat{\rho}(2))$$

Time series regression

There are a number of problems that may arise if serial correlation is ignored.

1. The estimated regression coefficients will still be unbiased, but no longer minimum variance.
2. Estimates of σ^2 will be biased.
3. The variance of the estimate of β will be underestimated and resulting t-statistics will be inflated.
4. Tests using the t and F distributions may not be applicable.

Time series regression

- When working with time series data, we typically use the index t to indicate the temporal ordering of observations.
- Throughout, we assume observations are measured at equally spaced time periods.
- A simple linear regression model for time series data is given by:

$$y_t = \beta_0 + \beta_1 X_t + \epsilon_t \quad t = 1, \ldots n$$

- We need to construct a model for ϵ_t that can account for autocorrelation.

Time series regression

- Commonly used models include autoregressive (AR), moving average (MA) and autoregressive-moving average (ARMA) models.
- Though the methods apply more generally, let us illustrate by assuming an AR(1) model, i.e.

$$X_t = \phi X_{t-1} + Z_t$$

where $Z_t \sim WN(0, \sigma^2)$ and $|\phi| < 1$.

AR(1) process

- The AR(1) process has the following properties:

$$E(X_t) = 0 \quad \forall t,$$

$$Var(X_t) = \frac{\sigma^2}{1-\phi^2} \quad \forall t$$

and

$$\begin{aligned} \gamma(h) &= \phi^{|h|}\gamma(0) \\ &= \frac{\sigma^2\phi^{|h|}}{1-\phi^2} \end{aligned}$$

Time series regression

- Consider the model: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim N(\mathbf{0}, \Sigma)$.
- Here we can write:

$$\Sigma = \frac{\sigma^2}{1-\phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^n \\ \phi & 1 & \phi & \cdots & \vdots \\ \phi^2 & \phi & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \phi \\ \phi^n & \phi^{n-1} & \cdots & \phi & 1 \end{pmatrix}$$

Time series regression

- When Σ is known we can estimate β using generalized least-squares.

$$\hat{\beta} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{y}.$$

- In general the form of the variance-covariance matrix is unknown, which means it has to be estimated.
- Estimating Σ depends on knowing β, and estimating β depends on knowing Σ.

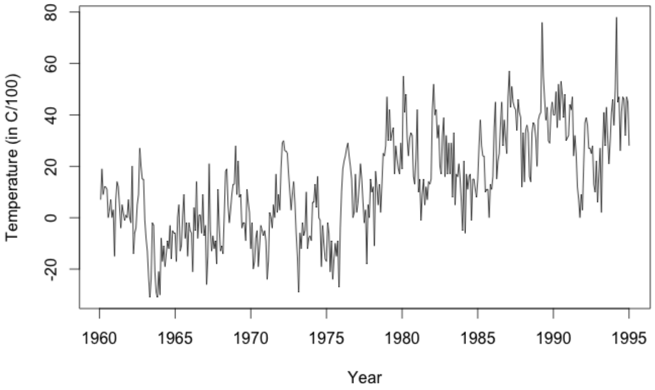
Cochrane-Orcutt Procedure

We need to use an iterative procedure, such as the Cochrane-Orcutt Procedure.

1. Assume that $\Sigma = \mathbf{I}\sigma^2$ and calculate the standard OLS solution.
2. Estimate the parameters of the time series model from the residuals.
3. Re-estimate the β values using the estimated covariance matrix from step 2.
4. Iterate until convergence.

Coding example

- The data consist of the monthly global mean temperature between 1961 and 1995.



Coding example

- Fit a model with a linear trend and a seasonal (monthly) effect, i.e.

$$y_t = \beta_0 + \beta_1 t + \beta_2 x_{2,t} + \cdots + \beta_{12} x_{12,t} + \epsilon_t$$

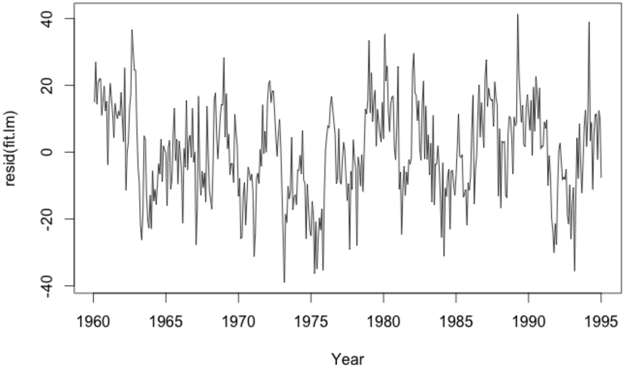
where $x_{i,t} = 1$ if t corresponds to month i, 0 otherwise.

```
> temp = scan('GlobalTemp')
> time = 1960+1:420/12
> season = factor(rep(1:12,35))
> fit.lm = lm(temp ~ time + season)
```

Coding example

- Study the residuals.

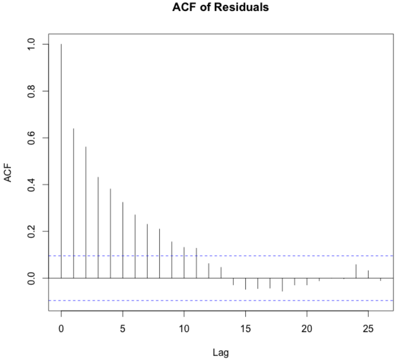
```
> plot(time, resid(fit.lm),xlab='Year', type="l")
```

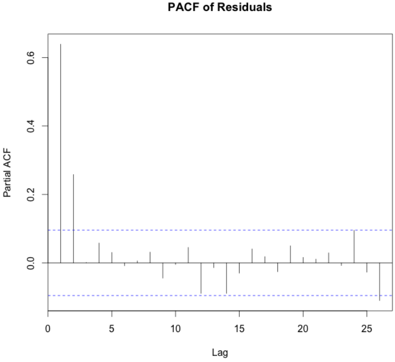


Coding example

- Study the ACF and PACF.

```
> acf(resid(fit.lm), main="ACF of Residuals")
> pacf(resid(fit.lm), main="PACF of Residuals")
```





Coding example

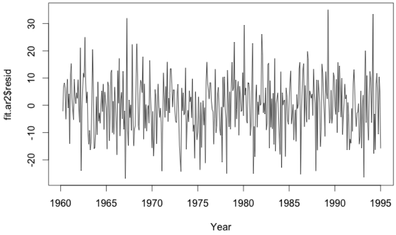
- Find the best fitting AR(p) model based on the residuals.

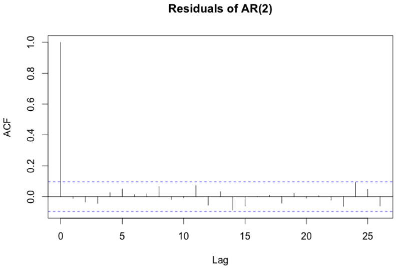
```
> fit.ar2 <- ar.yw(resid(fit.lm))
> fit.ar2

Call:
ar.yw.default(x = resid(fit.lm))

Coefficients:
     1       2
0.4654  0.2689

Order selected 2  sigma^2 estimated as  113.5
```

Coding example

- Study the residuals after fitting the AR(2) model.

```
> plot(time, fit.ar2$resid,xlab='Year', type="l")
> acf(fit.ar2$resid[3:420], main="Residuals of AR(2)")
```





Coding example

- Fit the model with an AR(2) error process.

```
> library(nlme)
> corStruct <- corARMA(p=2)
> fit.gls <- gls(temp~time+season, corr=corStruct)
> fit.gls
Generalized least squares fit by REML
  Model: temp ~ time + season
  Data: NULL
  Log-restricted-likelihood: -1569.99

Coefficients:
(Intercept)        time     season2     season3     season4     season5     season6
 -7.4439206   1.3466197  -0.1810691   2.3222767  -1.2313060  -2.2017252  -3.5726379
    season7     season8     season9    season10    season11    season12
 -3.6527144  -5.0699875  -5.4814065  -5.1935519  -5.0782568  -4.1397934

Correlation Structure: ARMA(2,0)
 Formula: ~1
 Parameter estimate(s):
     Phi1      Phi2
0.4663900 0.2781889
Degrees of freedom: 420 total; 407 residual
Residual standard error: 14.6746
```

Coding example

- Compare models with and without aurocorrelation model.

```
> coef(fit.lm)["time"]
    time
1.374621
> confint(fit.lm, "time")
        2.5 %   97.5 %
time 1.236521 1.512721

> coef(fit.gls)["time"]
   time
1.34662
> confint(fit.gls, "time")
         2.5 %   97.5 %
time 0.9589878 1.734252
```