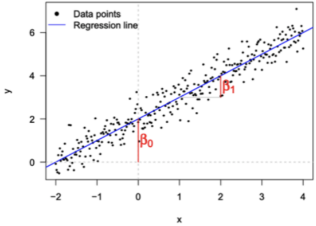
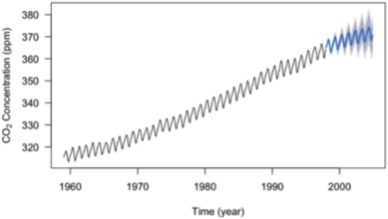
Lecture 15
Course Review

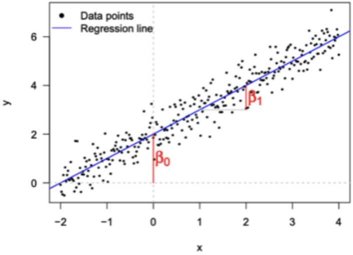
MATH 4070: Regression and Time-Series Analysis





Whitney Huang
Clemson University

Simple Linear Regression



$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2),$$

where

- β_0: intercept
- β_1: slope
- ε: random error

- **Parameter estimation**: Ordinary least squares (OLS)
- **Residual analysis**: For checking model assumptions
- **Statistical inferences**: Confidence/Predcition Interval; Hypothesis Testing
- **ANOVA**: To partition the total variability into regression and residual sums of squares

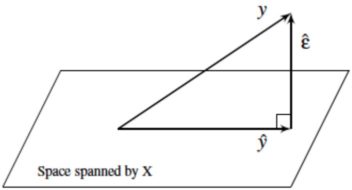
Multiple Linear Regression (MLR)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{p-1} X_{p-1} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2),$$

Matrix Notation:

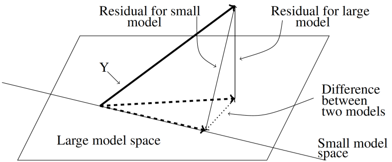
- **Data model**: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- **OLS**: $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$
- **Fitted values**: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$
  $= \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$
- **Residuals**: $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$
  $= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$
- **MSE**: $\frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-p}$

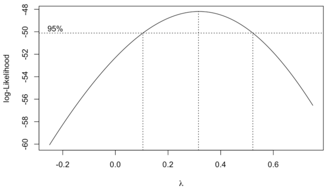
Geometric Representation:
Project the data vector $\boldsymbol{y}$ onto the (linear) model space



School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

MLR Additional Topics





- General Linear F-Test provides a unifying framework for hypothesis tests via full model vs. reduced model
- Multicollinearity, quantified via `VIF`, and its implications for MLR
- Model/variable selection can be done via some criterion-based methods (e.g., `AIC`) to balance **bias** and **variance**
- Box-Cox Transformation can be used to transform the response in order to alleviate model violations

Time Series Analysis Workflow

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

- **Plot the time series**

  $$Y_t = \mu_t + s_t + \eta_t$$

  Look for trends, seasonal components, step changes, and outliers.

- **Transform the data** so that the residuals are (approximately) stationary.
  - Apply nonlinear transformations (e.g., $\log$, $\sqrt{\cdot}$) to stabilize variance.
  - Use modeling (or differencing) to estimate (or remove) μ_t.
  - Use modeling (or differencing) to estimate (or remove) s_t.

- Identify potential (S)ARMA models for residuals and perform model fitting, selection, and diagnostics.

Stationarity

$\{Y_t\}$ is strictly stationary if, for all k, $t_1, \cdots, t_k$, $y_1, \cdots, y_k$ and h,

$$\mathbb{P}(Y_{t_1} \le y_1, \cdots, Y_{t_k} \le y_k) = \mathbb{P}(Y_{t_1+h} \le y_1, \cdots, Y_{t_k+h} \le y_k).$$

i.e., shifting the time axis does nor affect the joint distribution

We consider second-order properties only: $\{Y_t\}$ is stationary if its mean function and autocovariance function satisfy

$$\mu_t = \mathbb{E}[Y_t] = \mu,$$
$$\gamma(s,t) = \mathrm{Cov}(Y_s, Y_t) = \gamma(s-t).$$

Stationarity assumption $\Rightarrow$ consistent statistical properties over time $\Rightarrow$ enabling replication and allowing statistical modeling

ACF and Sample ACF

The autocorrelation function (ACF) is

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \mathrm{Cor}(Y_{t+h}, Y_t)$$

For observations $y_1, \cdots, y_n$ of a time series, the sample mean is

$$\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

The sample autocovariance function (ACVF) is

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-|h|}\left(y_{t+|h|} - \bar{y}\right)\left(y_t - \bar{y}\right), \quad \text{for } -n < h < n.$$

The sample autocorrelation function is

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

Linear Processes

Linear process is an important class of stationary time series:

$$Y_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j},$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$, and $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

- $\Rightarrow$ A linear time invariant filtering of $\{Z_t\}$ with coefficients $\{\psi_j\}$ that do not depend on time
- **Theorem**: Suppose $\{Z_t\}$ is a zero mean stationary series with ACVF $\gamma_Z(\cdot)$. Then $\{Y_t\}$ is a zero mean stationary process with ACVF

$$\gamma_Y(h) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \psi_j \psi_k \gamma_Z(j - k + h)$$

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

Causality and Inveribility

A linear porcess $\{Y_t\}$ is causal if there is a

$$\psi(B) = \psi_0 + \psi_1 B + \psi_2 B^2 + \cdots$$

with

$$\sum_{j=0}^{\infty} |\psi_j| < \infty \text{ and } Y_t = \psi(B) Z_t.$$

> All roots of the AR characteristic equation > 1 in modulus

A linear process $\{Y_t\}$ is invertible if there is a

$$\pi(B) = \pi_0 + \pi_1 B + \pi_2 B^2 + \cdots$$

with

$$\sum_{j=0}^{\infty} |\pi_j| < \infty \text{ and } Z_t = \pi(B) Y_t.$$

> All roots of the MA characteristic equation > 1 in modulus

Autoregressive Moving Average Models (ARMA)

An ARMA(p, q) process $\{Y_t\}$ is a stationary process that satisfies

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$. Also, $\phi_p, \theta_q \neq 0$ and $\phi(z)$ and $\theta(z)$ have no common factors

Properties:

- A unique stationary solution exists if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0 \Rightarrow |z| \neq 1.$$

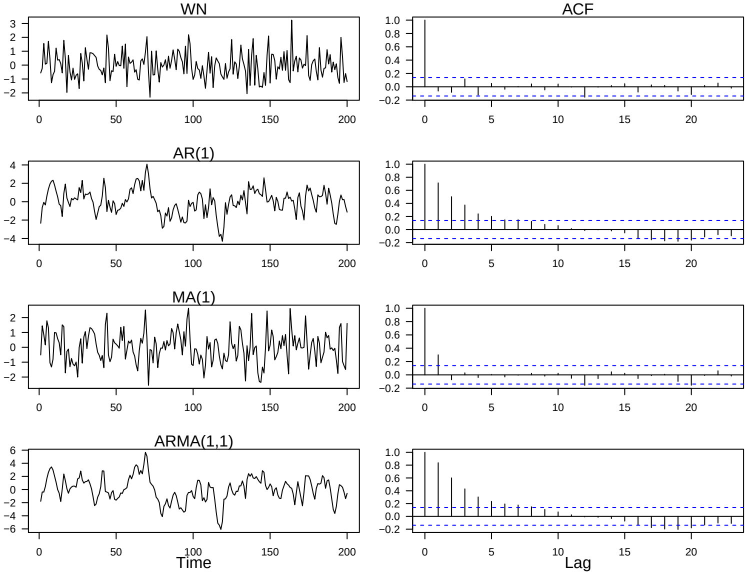
- This ARMA(p, q) process is causal if and only if

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p = 0 \Rightarrow |z| > 1.$$

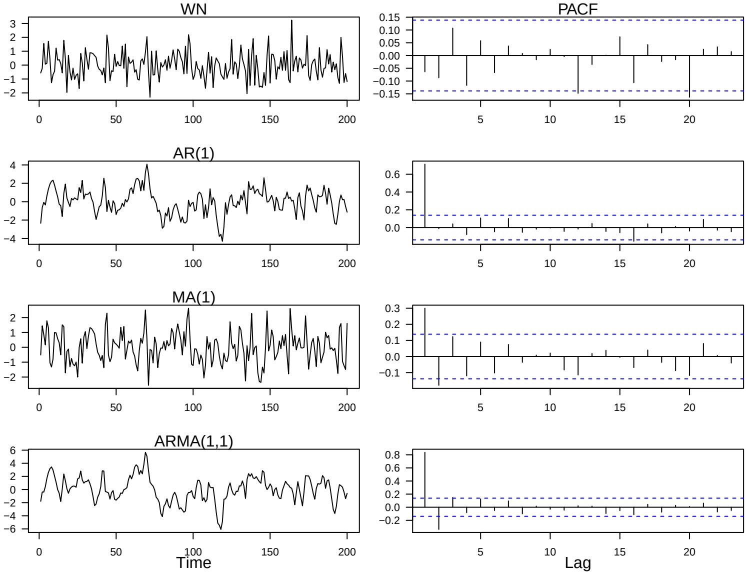
- It is invertible if and only if

$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q = 0 \Rightarrow |z| > 1.$$

ACF Plots



PACF Plots



Identification of ARMA Models using ACF/PACF Plots

Use the ACF and PACF together to identify candidate models. The following table gives some rough guidelines.

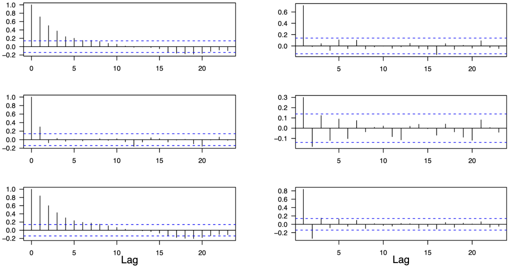
	ACF	PACF
AR(p)	Tails off	Cuts off after lag p
MA(q)	Cuts off after lag q	Tails off
ARMA(p, q)	Tails off	Tails off

ACF PACF

AR(1)

MA(1)

ARMA(1, 1)



Model Diagnostics: Ljung-Box Test

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

We wish to test:

$H_0 : \{e_1, e_2, \cdots, e_T\}$ is an i.i.d. noise sequence $\Rightarrow$ model adequate
$H_1 : H_0$ is false $\Rightarrow$ model not good,

where $\{e_t\}$ are the residuals after fitting a model to $\{\eta_t\}$

Test statistic:

$$Q_{LB} = T(T-2) \sum_{h=1}^{\text{lag}} \frac{\hat{\rho}^2_{\hat{\mathbf{e}}}(h)}{T-h} \overset{H_0}{\approx} \chi^2_k,$$

where T is the sample size, $\hat{\rho}_{\hat{\mathbf{e}}}(h)$ is the sample ACF at lag h, applied to the residuals of a fitted ARIMA model. The degrees of freedom $k = \text{Lag} - p - q$.

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

Linear Prediction

Given $Y_1, Y_2, \cdots, Y_n$, the best linear predictor $Y_{n+h}^n = \alpha_0 + \sum_{i=1}^n \alpha_i Y_i$ of Y_{n+h} satisfies the prediction equations:

$$\mathbb{E}[Y_{n+h} - Y_{n+h}^n] = 0$$
$$\mathbb{E}\left[(Y_{n+h} - Y_{n+h}^n)Y_i\right] = 0 \quad \text{for } i = 1, \cdots, n.$$

One-step-ahead linear prediction

$$Y_{n+1}^n = \phi_{n1} Y_n + \phi_{n2} Y_{n-1} + \cdots + \phi_{nn} Y_1$$
$$\Gamma_n \phi_n = \gamma_n, \quad P_{n+1}^n = \mathbb{E}(Y_{n+1} - Y_{n+1}^n)^2 = \gamma(0) - \gamma_n^T \Gamma_n^{-1} \gamma_n,$$

with

$$\Gamma_n = \begin{bmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(n-2) \\ \vdots & \cdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \cdots & \gamma(0) \end{bmatrix},$$

where

$$\phi_n = (\phi_{n1}, \phi_{n2}, \cdots, \phi_{nn})^T,$$

and

$$\gamma_n = (\gamma(1), \gamma(n), \cdots, \gamma(n))^T.$$

ARMA Parameter Estimation

Method of moments: choose parameters for which the moments are equal to the empirical moments. One choose ϕ such that $\gamma = \hat{\gamma}$.

Yule-Walker equations for $\hat{\phi}$: $\begin{cases} \hat{\Gamma}_p \hat{\phi} = \hat{\gamma}_p, \\ \hat{\sigma}^2 = \hat{\gamma}(0) - \hat{\phi}^T \hat{\gamma}_p. \end{cases}$

Maximum Likelihood Estimation: Suppose that $Y_1, \cdots, Y_n$ is drawn from a zero mean Gaussian ARMA(p, q) process. The likelihood of parameters $\phi \in \mathbb{R}^p$ and $\theta \in \mathbb{R}^q$, $\sigma^2 \in \mathbb{R}_+$ is defined as the joint density of $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$:

$$L(\phi, \theta, \sigma^2) = \frac{1}{(2\pi)^{n/2} |\Gamma_n|^{1/2}} \exp\left(-\frac{1}{2} \boldsymbol{Y}^T \Gamma_n^{-1} \boldsymbol{Y}\right).$$

The maximum likelihood estimator (MLE) of ϕ, θ, σ^2 maximizes this quantity.

ARIMA(p, d, q) and Seasonal ARIMA Models

For $p, d, q \geq 0$, we say that a time series Y_t is an ARIMA(p, d, q) process if

$$X_t = \nabla^d Y_t = (1 - B)^d Y_t$$

is ARMA(p, q). We can write

$$\phi(B)(1 - B)^d Y_t = \theta(B) Z_t.$$

For $p, q, P, Q \geq 0$, $s, d, D > 0$, we say a time series $\{Y_t\}$ is a seasonal ARIMA model (ARIMA(p, d, q) $\times$ $(P, D, Q)_s$) if

$$\Phi(B^s)\phi(B) \nabla_s^D \nabla^d Y_t = \Theta(B^s)\theta(B) Z_t,$$

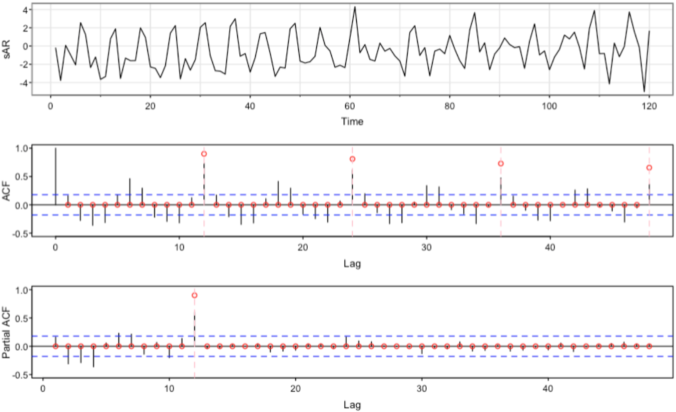
where the seasonal difference operator of order D is defined by

$$\nabla_s^D Y_t = (1 - B^s)^D Y_t.$$

An Example of a Seasonal AR Model

$$Y_t = 0.9Y_{t-12} + Z_t,$$

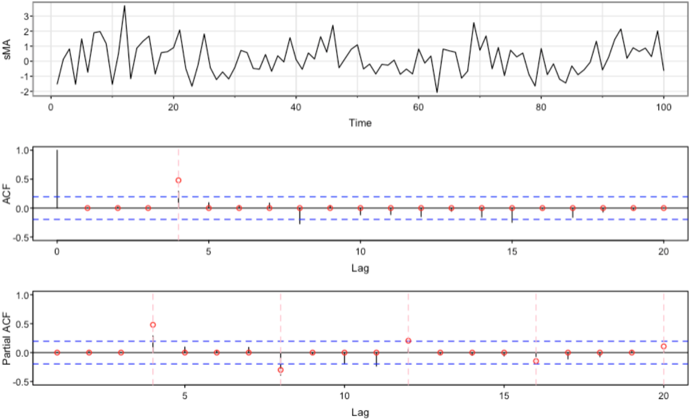
$$\Rightarrow p = q = d = D = Q = 0,\ P = 1,\ \Phi_1 = 0.9, s = 12.$$



An Example of a Seasonal MA Model

$$Y_t = Z_t + 0.75Z_{t-4},$$

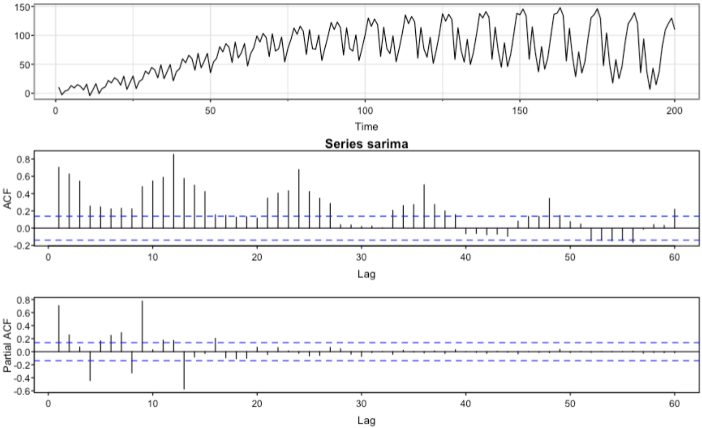
$$\Rightarrow p = q = d = D = P = 0,\ Q = 1,\ \Theta_1 = 0.75, s = 4.$$



Example of a SARIMA Model

$$(1-B)(1-B^{12})X_t = Y_t$$
$$(1+0.25B)(1-0.9B^{12})Y_t = (1+0.75B^{12})Z_t$$

$$\Rightarrow p = P = Q = d = D = 1,\ \phi = -0.25, \Phi = 0.9, \Theta_1 = 0.75, s = 12.$$



Generalized Least Squares Regression

When dealing with time series the errors $\{\eta_t\}$ are typically correlated in time

- Assuming the errors $\{\eta_t\}$ are a stationary Gaussian process, consider the model

  $$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta},$$

  where $\boldsymbol{\eta}$ has a multivariate normal distribution, i.e., $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \Sigma)$

- The generalized least squares (GLS) estimate of $\boldsymbol{\beta}$ is

  $$\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = \left(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{Y},$$

  with

  $$\hat{\sigma}^2 = \frac{\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)^T\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}\right)}{n - (p+1)}$$

School of MATHEMATICAL AND STATISTICAL SCIENCES Clemson University

Cross-Autocorrelation, Spurious Correlation, Prewhitening

The cross-covariance function of $\{Y_t\}$ and $\{X_t\}$ is

$$\gamma_{XY}(h) = \mathbb{E}\left[(X_{t+h} - \mu_X)(Y_t - \mu_Y)\right],$$

and the cross-correlation function (CCF) is

$$\rho_{XY}(h) = \frac{\gamma_{XY}(h)}{\sqrt{\gamma_X(0)\gamma_Y(0)}}.$$

CCF measures the correlation between two time series at different lags and helps detect lead-lag relationships

- **Spurious Correlation**: Misleading links caused by shared trends, seasonality, or confounders, often in non-stationary or autocorrelated data

- **Prewhitening**: Filtering out autocorrelation from one of the series to enable valid cross-correlation and reduce spurious results