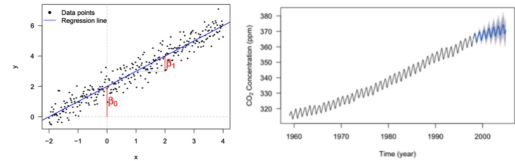


Lecture 15


Course Review

MATH 4070: Regression and Time-Series Analysis



Whitney Huang
Clemson University

Course Review

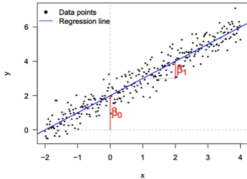


SCHOOL OF
MATHEMATICAL AND
STATISTICAL SCIENCES
CLEMSON UNIVERSITY

15.1

Notes

Simple Linear Regression




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

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/Prediction Interval; Hypothesis Testing
- **ANOVA:** To partition the total variability into regression and residual sums of squares

Course Review



SCHOOL OF
MATHEMATICAL AND
STATISTICAL SCIENCES
CLEMSON UNIVERSITY

15.2

Notes

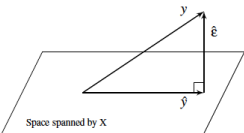
Multiple Linear Regression (MLR)

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


Matrix Notation:

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

Geometric Representation:
Project the data vector y onto the (linear) model space



Course Review

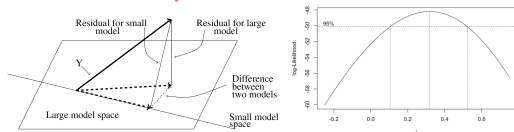


SCHOOL OF
MATHEMATICAL AND
STATISTICAL SCIENCES
CLEMSON UNIVERSITY

15.3

Notes

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

Course Review



15.4

Notes

Time Series Analysis Workflow

- **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**.

Course Review



15.5

Notes

Stationarity

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

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

i.e., shifting the time axis does not affect the joint distribution

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

$$\begin{aligned} \mu_t &= \mathbb{E}[Y_t] = \mu, \\ \gamma(s, t) &= \text{Cov}(Y_s, Y_t) = \gamma(s - t). \end{aligned}$$

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

Course Review



15.6

Notes

ACF and Sample ACF

The autocorrelation function (ACF) is

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

For observations y_1, \dots, 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

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

The sample autocorrelation function is

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

Course Review



15.7

Notes

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

- ⇒ 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

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

Course Review



15.8

Notes

Causality and Invertibility

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

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

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

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

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

Course Review



15.9

Notes

Autoregressive Moving Average Models (ARMA)

An ARMA(p, q) process {Yt} is a stationary process that satisfies

Yt - φ1Yt-1 - ... - φpYt-p = Zt + θ1Zt-1 + ... + θqZt-q,

where {Zt} ~ WN(0, σ²). Also, φp, θq ≠ 0 and φ(z) and θ(z) have no common factors

Properties:

- A unique stationary solution exists if and only if

φ(z) = 1 - φ1z - ... - φpz^p = 0 ⇒ |z| ≠ 1.

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

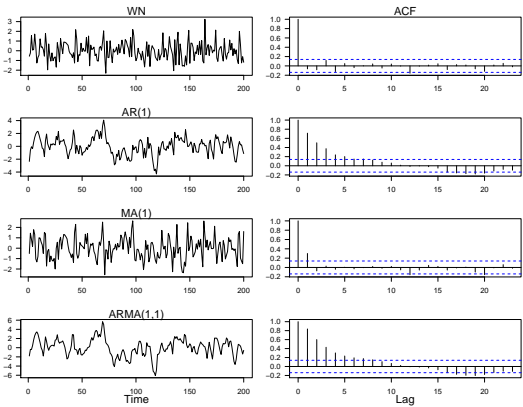
φ(z) = 1 - φ1z - ... - φpz^p = 0 ⇒ |z| > 1.

- It is invertible if and only if

θ(z) = 1 + θ1z + ... + θqz^q = 0 ⇒ |z| > 1.

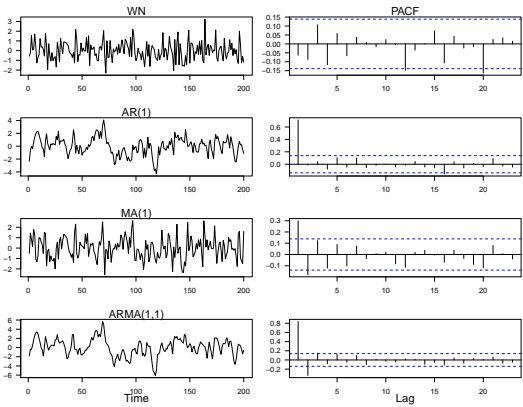
Notes

ACF Plots



Notes

PACF Plots

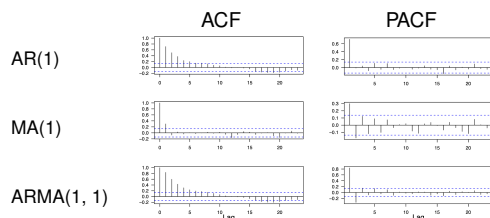


Notes

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



Course Review



15.13

Notes

Model Diagnostics: Ljung-Box Test

We wish to test:

$H_0 : \{e_1, e_2, \dots, e_T\}$ 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:

$$Q_{LB} = T(T-2) \sum_{h=1}^{\text{lag}} \frac{\hat{\rho}_e^2(h)}{T-h} \stackrel{H_0}{\approx} \chi_k^2,$$

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

Course Review



15.14

Notes

Linear Prediction

Given Y_1, Y_2, \dots, 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:

$$\mathbb{E}[Y_{n+h} - Y_{n+h}^n] = 0$$

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

One-step-ahead linear prediction

$$Y_{n+1}^n = \phi_{n1}Y_n + \phi_{n2}Y_{n-1} + \dots + \phi_{nn}Y_1$$

$$\Gamma_n \phi_n = \gamma_n, \quad P_{n+1}^n = \mathbb{E}(Y_{n+1} - Y_{n+1}^n)^2 = \gamma(0) - \gamma_n^T \Gamma_n^{-1} \gamma_n,$$

with

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

where

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

and

$$\gamma_n = (\gamma(1), \gamma(2), \dots, \gamma(n))^T.$$

Course Review



15.15

Notes

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, \dots, 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 $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)$:

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

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



Notes

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

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

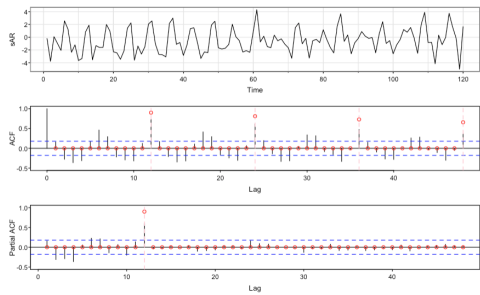


Notes

An Example of a Seasonal AR Model

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

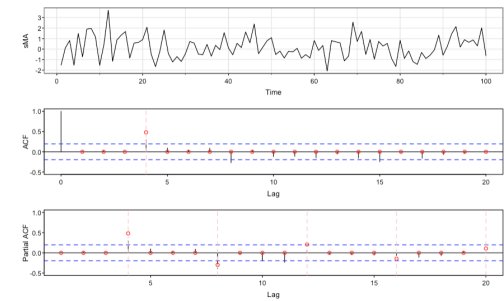
$$\Rightarrow p = q = d = D = Q = 0, P = 1, \Phi_1 = 0.9, s = 12.$$



Notes

An Example of a Seasonal MA Model

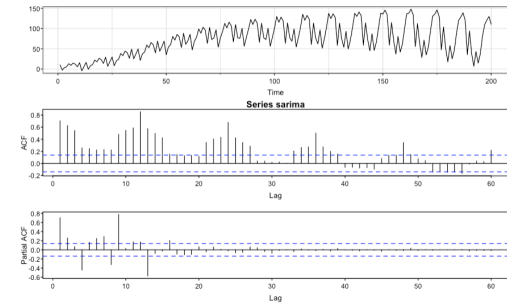
$$Y_t = Z_t + 0.75Z_{t-4},$$
$$\Rightarrow p = q = d = D = P = 0, Q = 1, \Theta_1 = 0.75, s = 4.$$



Notes

Example of a SARIMA Model

$$(1 - B)(1 - B^{12})X_t = Y_t$$
$$(1 + 0.25B)(1 - 0.9B^{12})Y_t = (1 + 0.75B^{12})Z_t$$
$$\Rightarrow p = P = Q = d = D = 1,$$
$$\phi = -0.25, \Phi = 0.9, \Theta_1 = 0.75, s = 12.$$



Notes

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

$$Y = X\beta + \eta,$$

where η has a multivariate normal distribution, i.e., $\eta \sim N(0, \Sigma)$

- The generalized least squares (GLS) estimate of β is

$$\hat{\beta}_{\text{GLS}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y,$$

with

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

Notes

Cross-Autocorrelation, Spurious Correlation, Prewhitening

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

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

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

Notes

Notes

Notes
