IE5202: Applied Forecasting Methods

Chen Nan

August 7, 2019

Contents

1 Syllabus

1.1 Module Information

Instructor: Dr Chen Nan, E1-05-20

Contact: Phone: 65167914
Email: isecn@nus.edu.sg

TA: Xie Jiaohong (xiejiaohong@u.nus.edu)

Office Hours: By appointment

Textbook: *Forecasting, Time Series, and Regression*, by Bowerman, O'Connell, and Koehler

References: *Linear Regression Analysis*, by George A. F. Seber, Alan J. Lee
Time Series Analysis, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel

Prerequisites: IE 5002, IE 6002, programming

Description: This module focuses on the theory and practice of forecasting methods. It discusses two major categories of forecasting problems, and corresponding techniques. Extensive hands on projects will be provided to solve real life problems.

1.2 Schedule

Aug 12, 2019	*Hari Raya Haji (Public Holiday)*
Aug 19, 2019	Module logistics, introduction, and reviews
Aug 26, 2019	Regression analysis
Sep 02, 2019	Model checking and diagnosis
Sep 09, 2019	Model evaluation & selection
Sep 16, 2019	Machine learning approaches
Sep 23, 2019	*Recess Week*
Sep 30, 2019	Seasonality, regression on time
Oct 07, 2019	Exponential smoothing
Oct 14, 2019	ARMA, SARMA
Oct 21, 2019	Neural networks for time series
Oct 28, 2019	*Deepavali (Public Holiday)*
Nov 04, 2019	Forecasting spatial data
Nov 11, 2019	TBD

1.3 Grading

Grading: Homework: 10%
Project 1: 30% ..... Due Sep 29, 2019
Project 2: 30% ..... Due Nov 17, 2019
Final Exam: 30% ..... 04 Dec 2019 9:00-11:00am

The projects are based on real problems and real data. The dataset and background information of the project will be provided. For each project, the submission should include the following items

- A report not more than 10 pages with 1.5 spacing (soft copies and hard copies), which documents the methods using, main findings, and interpretations. Codes and software printouts should NOT be included in the report.
- Complete codes used for the analysis, with reasonable details of comments (Soft copies only)
- Forecasting results on the test dataset in a "csv" file with a single column, as shown in the following example

```
A0001124H
10.31
8.5
20.1
...
11.5
```

2 Introduction

2.1 Definitions

Definition 2.1 *Predictions of future events and conditions are called* ***forecasts****, and the act of making such predictions is called* ***forecasting****.*

Forecasting is very important in many types of organizations since predictions of future events must be incorporated into the decision-making process. Examples include

- Government needs to forecast such things as air quality, water quality, unemployment rate, inflation rate, and welfare payment, etc.
- A university needs to forecast its enrollment, temperature, broken asset.
- Business firms need to forecast demands to plan sales and production strategy, to forecast interest rate for financial planning, to forecast number of workers required for human resource planning, and to forecast the quality of the product for process improvement and quality control.

To forecast events that will occur in the future, one must rely on information concerning events that have occurred in the past. Based on the type of information used, there are two categories of forecasting methods.

1. **Qualitative forecasting methods**: use the opinions of experts to subjectively predict future events. It is often required when historical data are either not available or scarce, or when changes in data pattern cannot be predicted on the basic of historical data. Commonly used qualitative methods include

   (a) Delphi Method: Use a panel of experts to produce predictions concerning a specific question such as when a new development will occur in a particular field. The panel members are kept physically separated. After the first questionnaire has been completed and sent, subsequent questionnaires are accompanied by information concerning the opinions of the group as a whole.

   (b) Technological comparisons: are used in predicting technological change. It determines a pattern of change in one area, called a primary trend, which the forecaster believes will result in new developments being made in some other area. A forecast of developments in the second area can then be made by monitoring developments in the first area.

   (c) Subjective curve fitting: The forecaster subjectively determines the form of the curve to be used, and a great deal of the expertise and judgment is required.

2. **Quantitative forecasting methods**: involves the analysis of historical data in an attempt to predict future values of a variable of interest. The methods often depend on the types of data available.

   Definition 2.2 ***Cross-sectional*** *data are values observed at one point in time; A* ***time series*** *is a chronological sequence of observations on a particular variable.*

   As a result, the Quantitative methods can be roughly classified as

   (a) Causal methods: involve the identification of other variables that are related to the variable to be predicted. It develop a statistical model that describes the relationship between these variables and the variable to be forecasted. For example, the sales of a product might be related to the price of the product, competitors' prices for similar products, advertising expenditures to promote the products.

   (b) Time series methods: make prediction of future values of a time series based solely on the basis of the past values of the time series. It tries to identify a pattern in the historical data, which is extrapolated in order to make a forecast. It is assumed that the pattern will continue in the future. For example, one predicts the temperature tomorrow based solely on the temperatures in the past days.

2.2 Deterministic vs Stochastic Relations

theoretical vs experimental

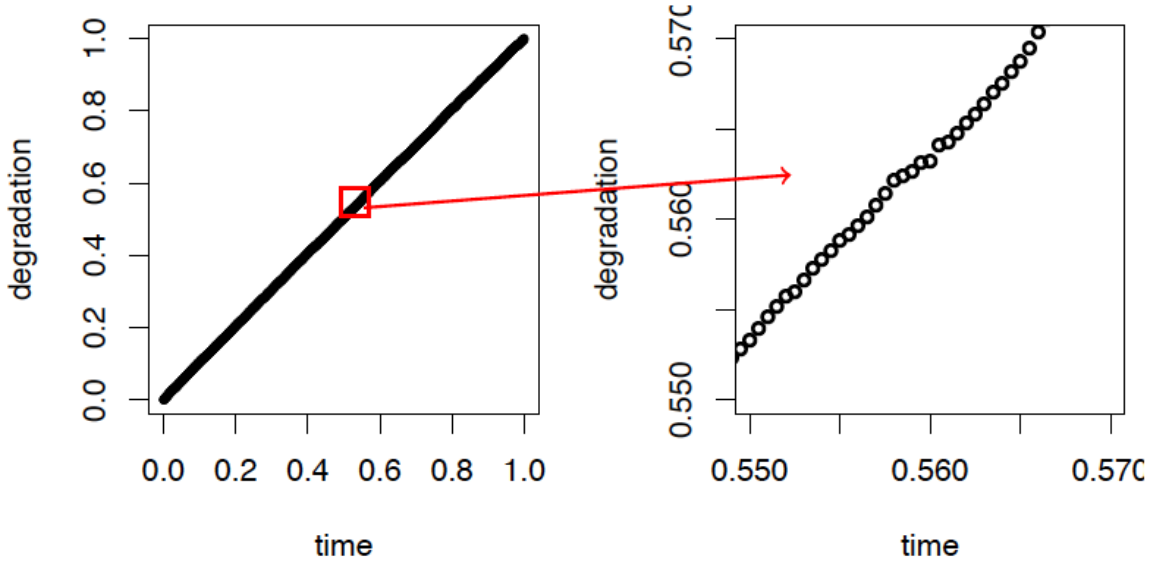
In this module, we only focus on *quantitative* methods for forecasting. Before proceeding, we must clarify the types of relationships we do not study in this module, namely, deterministic (or functional) relationships. Here are examples of a deterministic relationship.

- As you may know , the relationship between degrees Fahrenheit and degrees Celsius is known to be: $\text{Fahr} = 9 \times \text{Cels}/5 + 32$. If you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit exactly.
- Circumference $= \pi \times$ diameter
- Hooke's Law: $Y = \alpha + \beta X$, where Y is amount of stretch in a spring, and X is the applied weight.
- Ohm's Law: $I = V/r$, where V is the voltage applied, r is the resistance, and I is the current.
- Boyle's Law: For a constant temperature, $P = \alpha/V$, where P is pressure, α is a constant for each gas, and V is the volume of the gas.

For each of these deterministic relationships, the equation exactly describes the relationship between the two variables. This course does not examine deterministic relationships. Instead, we are interested in *statistical* or stochastic relationships, in which the relationship between the variables is not perfect. Some examples of statistical relationships might include:

- Height and weight: as height increases, you'd expect weight to increase, but not perfectly.
- Alcohol consumed and blood alcohol content: as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
- Vital lung capacity and pack-years of smoking: as amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
- Driving speed and gas mileage: as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

It is also noted that the boundary between deterministic and stochastic relationship might not be clear in some scenarios. For example, depending on the accuracy and precision requirement, Newton's laws in physics can be viewed as deterministic relations in some cases, but can only serve as approximation to theory of relativity in some other cases. It is also possible that the stochastic or random elements observed are simply due to some unknown variables in a deterministic relations.



2.3 Errors in Forecasting

Unfortunately, all (stochastic) forecasting involve some degree of uncertainty. We recognize this fact by including an irregular component in the description of the model. The presence of this irregular component, which represents *unexplained* or *unpredictable* fluctuations in the data, means that some error in forecasting must be expected.

The fact that forecasting techniques often produce predictions that are somewhat in error has a bearing on the form of the forecasts we require. Two types of forecasts are common in practice.

- **Point forecast**: a single number representing our "best" prediction of the actual value
- **Prediction interval forecast**: an interval of numbers that will contain the actual value with certain confidence (95%)

To evaluate the performance or accuracy of the forecasting methods, certain criteria shall be used. For point forecast, a natural way is to calculate the forecast error.

Definition 2.3 *The forecast error for a particular forecast $\hat{y}$ of a quantity of interest y is*

$$e = y - \hat{y}$$

If the forecast is accurate, the error is small stochastically. In general, e cannot be zero, and can be large even for a good forecasting method in "unlucky" cases. Therefore, it is important to measure the magnitude of the errors over time or over different samples to evaluate the forecasting method.

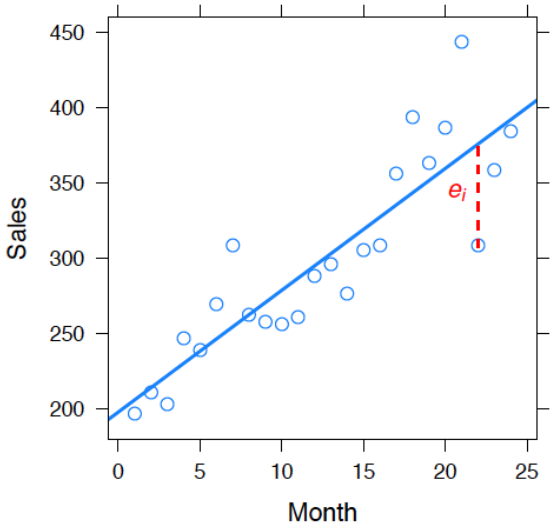
Definition 2.4 *The mean absolute deviation (MAD) of the forecasting is defined as*

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|.$$

The mean squared error (MSE) of the forecasting is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|^2.$$

Intuitively, MSE is influenced by large forecast errors more sensitively.



To compare the forecast on quantities of different scales, relative errors can be adopted by normalizing the error by the value to be forecasted.

Definition 2.5 *The mean* ***absolute percentage errors*** *(MAPE) of a method is defined as*

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|.$$

Different from point forecast, prediction interval forecast is often an interval of values. Its performance depends on two factors.

Definition 2.6 ***Coverage probability*** *is the proportion of the time that a confidence interval contains the true value of interest.* ***Length of the interval*** *is simply the difference in the two endpoints.*

Ideally, for a 95% prediction interval, the interval should have coverage probability 0.95, i.e., containing the true value 95% of the time. On the other hand, the length of the interval indicates the precision of the forecast. Given the same coverage probability, the shorter interval length the better.

3 Linear Regression

3.1 Overview

Galton was a pioneer in the application of statistical methods. In studying data on relative sizes of parents and their offspring in various species of plants and animals, he observed the following phenomenon: a larger-than-average parent tends to produce a larger-than-average child, but the child is likely to be less large than the parent in terms of its relative position within its own generation. Galton termed this phenomenon a regression toward mediocrity, which in modern terms is a *regression to the mean*.

Regression to the mean is an inescapable fact of life. Your children can be expected to be less exceptional (for better or worse) than you are. Your score on a final exam in a course can be expected to be less good (or bad) than your score on the midterm exam, relative to the rest of the class. The key word here is "expected". This does not mean it's certain that regression to the mean will occur, but it has a higher chance than not. For detailed account for this, please refer to http://www.socialresearchmethods.net/kb/regrmean.php.

Linear regression analysis is the most widely used of all statistical techniques: it is the study of linear, additive relationships between variables. Even though the *linear* condition seems restrictive, it has some practical justifications:

- linear relationships are the simplest non-trivial relationships;

- the "true" relationships between variables are often approximately linear, at least over a range of values;
- Even for some nonlinear relationships, we can often transform the variables in a way to linearize the relationships.

3.2 Simple linear regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. One variable, denoted x, is regarded as the *predictor, explanatory, or independent* variable. The other variable, denoted y, is regarded as the *response, outcome, or dependent* variable. We will use the **predictor** and **response** terms to refer to the variables encountered in this module.

Linear regression attempts to describe the nature of the association by constructing a "best-fitting" mathematical model. When using linear regression we assume that the variables are associated in a linear fashion, and we attempt to find that line that explains the association. Mathematically, it describes the relation between response Y and predictor X as

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{3.1}$$

where $\mathbb{E}\epsilon = 0$, and $\sigma^2 = \text{var}(\sigma) < \infty$. ϵ indicates the observation error, noise, or uncertainties that cannot be accounted for by the linear relation $\beta_0 + \beta_1 X$. Here "*linear*" is to quantify the relationship in parameters, which means the partial derivative w.r.t. $\boldsymbol{\beta}$ should be free of all parameters.

3.2.1 Least square estimation

The linear relation (linear model) (3.1) has three parameters $\beta_0, \beta_1, \sigma^2$. They have clear physical interpretations. However, in practice their values might not be available, and need to be estimated from observations.

Assume that we have n observations of (x_i, y_i). We want to find a straight line that "best" forecast (approximate) these n points. For any given values $\beta_0 = a, \beta_1 = b$, a natural point forecast of the response given predictor X is simply $a + bX$, the conditional expectation $\mathbb{E}(Y|X)$. Recall the commonly used forecasting error MSE is defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

The least square estimation of the parameters are defined as the values of β_0, β_1 that can minimize the MSE

$$\hat{\beta}_0, \hat{\beta}_1 = \arg\min_{a,b} \sum_{i=1}^{n}(y_i - a - bx_i)^2/n \tag{3.2}$$

By taking the derivatives with respect to a, b, we can get the analytical expression for $\hat{\beta}_0, \hat{\beta}_1$ as

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},\end{aligned} \tag{3.3}$$

where $\bar{x} = \sum_{i=1}^n x_i/n$ and $\bar{y} = \sum_{i=1}^n y_i/n$ are the sample average of x_i, y_i respectively. It is easy to show that $\mathbb{E}\hat{\beta}_0 = \beta_0$, and $\mathbb{E}\hat{\beta}_1 = \beta_1$, meaning they are *unbiased* estimators of the regression coefficients.

A natural way to estimate the variance σ^2 is by

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \tag{3.4}$$

where the term $\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ is often named as the sum of squared errors (SSE). The term $(n-2)$ is to make the estimation unbiased, $\mathbb{E}\hat{\sigma}^2 = \sigma^2$. $\hat{\sigma}^2$ quantifies the uncertainty of the regression line, and is related to goodness-of-fit as well.

Remark: There are a few interesting notes for the least square estimation in simple linear regression

- The estimated regression line $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ passes the point $(\bar{x}, \bar{y})$.
- The estimated slope $\hat{\beta}_1$ is closely related to the correlation coefficients between X and Y. Recall that the sample correlation is defined as

  $$\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

  Comparing it with $\hat{\beta}_1$, we can find that

  $$\hat{\beta}_1 = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \hat{\rho} = \frac{S_Y}{S_X} \hat{\rho},$$

  where S_Y^2, S_X^2 are the sample variance of Y, X respectively.
- The denominator $n-2$ in (3.4) is used to make $\hat{\sigma}^2$ unbiased, i.e., $\mathbb{E}\hat{\sigma}^2 = \sigma^2$.

3.2.2 Accuracy of the estimation

The aforementioned results do not require the specific form of the error distribution. However, to assess the accuracy of the estimation, and to construct the confidence interval, it is necessary to know the distribution of ϵ. A commonly used assumption is that ϵ_i follows normal distribution $\epsilon_i \sim N(0, \sigma^2)$, independently.

Based on the linear regression model, $Y = \beta_0 + \beta_1 X + \epsilon$, we can also conclude that the conditional distribution of $P(Y|X) \sim N(\beta_0 + \beta_1 X, \sigma^2)$. It is clear that the regression part $\beta_0 + \beta_1 X$ models the mean (expectation) of the response, and the error term quantifies the uncertainty in the prediction (modeling).

If y_i are independent, and the true relationship follows the model (3.1), it can be derived that the parameter estimation (3.3) follows normal distribution with

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right), \quad \hat{\beta}_0 \sim N\left(\beta_0, \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\sigma^2\right). \tag{3.5}$$

It shows the least square method can estimate the true parameters "on average". However, given finite number of samples, there exists uncertainty in estimating the true parameters. The magnitude of uncertainty depends on two factors:

- The sample size n used in the estimation
- The scatters of x_i. The more dispersed of x_i, the better estimation.

Similarly, we can find the distribution of $\hat{\sigma}^2$ with normal assumption. Using the formula in (3.4), it can be derived that

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-2} \cdot \chi^2_{n-2}. \tag{3.6}$$

As a result, we have $\mathbb{E}\hat{\sigma}^2 = \sigma^2$.

The distribution of estimated parameters also allow us to construct the confidence intervals for such estimates. From (3.5) and (3.6), we can find that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{var}(\hat{\beta}_1)}} \sim t_{n-2}, \quad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\text{var}(\hat{\beta}_0)}} \sim t_{n-2}.$$

As a result, the $1-\alpha$ level confidence interval for both parameters are

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad \hat{\beta}_0 \pm t_{n-2,\alpha/2} \cdot \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{3.7}$$

where $t_{n-2,\alpha/2}$ is the upper $\alpha/2$ quantile of the t_{n-2} distribution.

Remark: Some additional notes on the estimation accuracy.

- We can construct the confidence interval for σ^2 as well based on (3.6).
- Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated based on the same set of data, they are not independent. The covariance between them is crucial in making interval forecasting.

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2. \tag{3.8}$$

- When the number of samples is large enough ($n \to \infty$), we can expect $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2$ all *converge* to the true values.

3.2.3 Point forecasting

Given the estimated parameters based on least square methods, we can make point forecast given any value of the predictor X. In fact, the "best" prediction is the conditional mean. Given $X = x^*$, the point forecast $y*$ is simply

$$y^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

Substituting the formulas in (3.3), we have

$$y^* = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x^* = \bar{y} + \hat{\beta}_1 (x^* - \bar{x}).$$

Essentially, the forecasted value of the response, in terms of deviation from the mean, is proportional to the deviation of the predictor value from its mean.

Using the relationship between $\hat{\beta}_1$ and the correlation $\hat{\rho}$, we can also write the forecasting as

$$\frac{y^* - \bar{y}}{S_Y} = \hat{\rho} \frac{x^* - \bar{x}}{S_X}, \tag{3.9}$$

where $(y^* - \bar{y})/S_Y$ and $(x^* - \bar{x})/S_X$ can be viewed as "standardized" value. Since $\hat{\rho}$ is always smaller than 1, this also explains the "regression to mean" technically. In particular, our prediction for standardized y^* is typically smaller in absolute value than our observed value for standardized x^*. That is, the prediction for Y is always closer to its own mean, in units of its own standard deviation, than X was observed to be, which is Galton's phenomenon of regression to the mean. The perfect positive correlation ($\rho = 1$) or perfect negative correlation ($\rho = -1$) is only obtained if one variable is an exact linear function of the other, without error, i.e., $Y = \beta_0 + \beta_1 X$. In this case, the relationship between X and Y becomes deterministic rather than stochastic.

3.2.4 Interval Forecasting

When we use the estimated parameters to make forecasting, we need to consider the uncertainty in the estimated parameters. In particular, when the predictor has value $X = x^*$, the conditional mean of the response $y^* = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^*$, as shown in Section 3.2.3. Using the distributional information on $\hat{\beta}$, we can also find the distribution of y^*. Because both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed (3.5), we can derive

$$P(y^*|x^*) \sim N\left(\beta_0 + \beta_1 x^*, \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \sigma^2\right). \tag{3.10}$$

It is clear that the parameter uncertainty propagates to the forecast uncertainty. It still has the true mean "on average", but with additional variation as the variance component in the normal

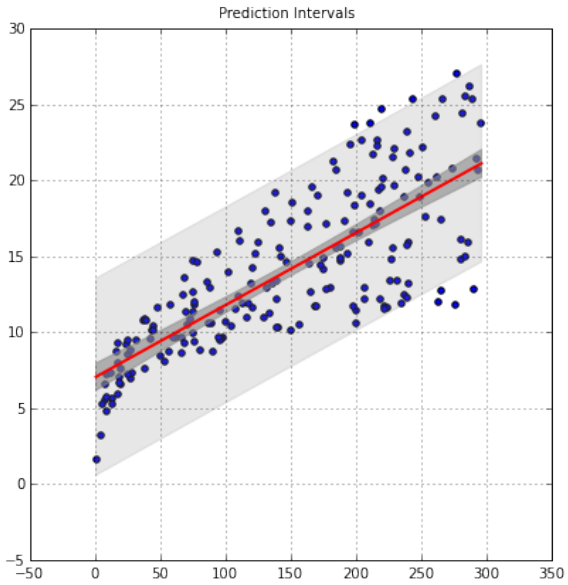
distribution (3.11). The form of the variance also suggest that the forecasting accuracy (or equivalently the magnitude of the variance) depends on the following factor: (a) the sample size n used in estimation; (b) the scatters of the observations x_i; (c) the distance between the forecast point x^* and the data center $\bar{x}$.

From (3.11), we can construct the prediction interval forecast at $1-\alpha$ confidence level:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}. \tag{3.11}$$

It is noted that the prediction interval is for the mean value of the response at x^*. It is different from the prediction interval of the response, which includes another error term ϵ. Consequently, the prediction interval for the response is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}. \tag{3.12}$$

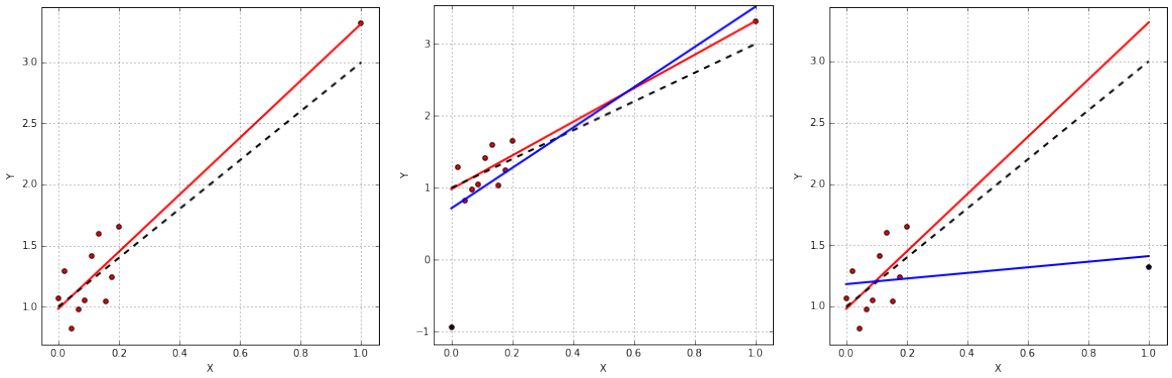


3.2.5 Other notable notes

1. **Model assumptions**: most results discussed above rely on some important assumptions of the data. In addition to the linear form of the mean $\beta_0 + \beta_1 X$, the error term ϵ_i must be independent, and normally distributed, with the *same* variance. These assumptions must be checked after the model estimation (discussed later). In many problems, these assumptions might be severely violated. In these cases, the model needs to be revised before reaching meaningful results.

2. **Transformation**: The *linear* constraint does not imply the relationship can only be a straight line. In fact, different transformations on x can be used (e.g., x^2, $\ln(x)$, $\sqrt{x}$). The results discussed above still hold with transformed variables. The transformation can be inspired from data feature, or guided by first-principles.

3. **Outliers**: The least square method minimizes the sum of squared error, as a result it is sensitive to outliers. A single outlier can drive the estimated parameters far from its true values. Therefore, it is important to recognize outliers, especially to differ between outliers and normal large values.



   - **Leverage**: is a measure of how far away the predictor values of an observation are from those of the other observations
   - **Outliers** are values that cause surprise in relation to the majority
   - **Influential** observations have a relatively large effect on the regression model's predictions

4. **Interpretation**: When making forecasting or interpreting the results, it is important to understand the limits of the model. For example, a linear relationship might be satisfactory in a range of temperature, but loses its validity when extended to the cases outside the range. Special attention is required when making forecasts far from the center $\bar{x}$ of historical data.

3.3 Multiple linear regression

Multiple linear regression attempts to model the relationship between *two or more* explanatory variables and a response variable by fitting a linear equation to observed data. If there are p predictors (or explanatory variables) $x_1, x_2, \cdots, x_p$, for every combination of the predictor variables $x_{i1}, x_{i2}, \cdots, x_{ip}$, it is associated with a value of the response variable y_i. Similar to the simple regression, the multiple linear regression define the regression model as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon.$$

This model describes how the mean response $\mathbb{E}(Y)$ changes with the explanatory variables. The observed values for Y_i vary about their means and are assumed to have the same standard deviation σ. The parameters of the model include $\beta_0, \beta_1, \cdots, \beta_p, \sigma^2$.

For concise representation, we often write the model in matrix/vector form. Define $\boldsymbol{\beta} = [\beta_0, \beta_1, \cdots, \beta_p]$, and $\mathbf{x}_i = [1, x_{i1}, \cdots, x_{ip}]$, then we have

$$Y_i = \mathbf{x}_i \cdot \boldsymbol{\beta} + \epsilon_i, \tag{3.13}$$

with $\mathbb{E}(\epsilon_i) = 0, \text{var}(\epsilon_i) = \sigma^2$ again. Throughout the handout, we will use bold symbols to represent vectors or matrices. Combine all observations together, we have:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

or,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.14}$$

3.3.1 Why multiple regression?

Simple linear regression only allows a single predictor in modeling the response. This might be too restrictive in many cases. The following examples illustrate the need for multiple linear regression.

I. Complex relation with a single predictor variable. Even when the response is related to a single predictor variable, the relation might be more complex than a straight line. Consider a commonly used model in practice, the polynomial regression

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$

The response changes with x in a quadratic way. Only in special cases ($\beta_1 = 0$ or $\beta_2 = 0$), can we use simple linear regression to estimate the parameters.

II. Relation with qualitative predictor. The simple linear regression implies the predictor is a quantitative (continuous) variable, so that the multiplication and addition have meaning. However, we often encounter qualitative variables, such as gender, color, race, etc. They are often called attribute variables or *factors*. Since they do not have natural ordering, numerical operations on the variable lose the validity.

Taking Race with four values (Chinese, Malay, Indian, Caucasian) as an example. Instead of doing regression directly on this variable, some dummy variables can be created.

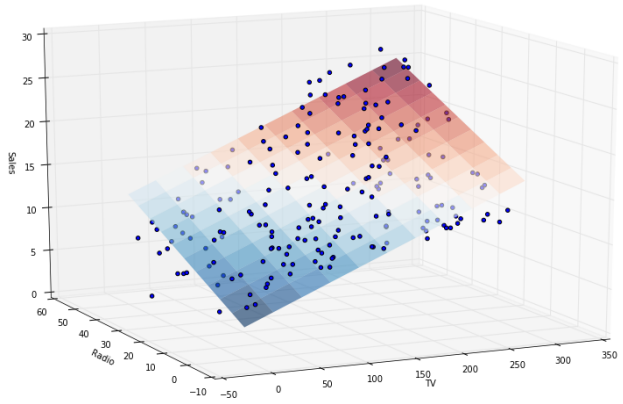
Original	Transformed			
X	Chinese(R1)	Malay(R2)	Indian(R3)	Caucasian(R4)
Chinese	1	0	0	0
Caucasian	0	0	0	1
Indian	0	0	1	0
Indian	0	0	1	0
Chinese	1	0	0	0
Malay	0	1	0	0

As a result, instead of having $Y = \beta_0 + \beta_1 X$, which is not meaningful here, we can have $Y = \beta_1 R1 + \beta_2 R2 + \beta_3 R3 + \beta_4 R4$ using the dummy variables. The coefficients have clear interpretation in this case. Again, even with a single predictor (Race in this example), we still need multiple linear regression. Another example is salary in IT industry (many years ago) versus the education level.

	S	X	E	M	X1	X2	X3
0	13876	1	1	1	1.	0.	0.
1	11608	1	3	0	1.	0.	1.
2	18701	1	3	1	1.	0.	1.
3	11283	1	2	0	1.	1.	0.
4	11767	1	3	0	1.	0.	1.

Here "S" denotes for salary, "E" denotes for education level, "M" denotes for management or non-management, "X" denotes for experience.

III. Multiple variables with non-separable effects. More commonly, a response variable is often influenced by multiple predictors. It is generally not sensible to quantify their influence one by one using simple linear regression. In addition, sometimes explicit *interaction* between two variables are required. The interactions will be discussed in more details later. An example here is the joint effects of TV and Radio on sales: $Y = 2.94 + 0.046 * TV + 0.19 * Radio - 0.001 * Newspaper$



3.3.2 Interactions between variables

There are two implicit assumptions when formulating the multiple linear regression: (1) Effects of different predictors are additive. (2) If x_1 changes Δx_1, the mean response always changes $\beta_1 \Delta x_1$, regardless other predictors. In statistics, an *interaction* may arise when considering the relationship among three or more variables, and describes a situation in which the simultaneous influence of two variables on a third is not *additive*.

The presence of interactions can have important implications for the interpretation of statistical models. If two variables of interest interact, the relationship between each of the interacting variables and the response variable depends on the value of the other interacting variable. In practice, this makes it more difficult to predict the consequences of changing the value of a variable, particularly if the variables it interacts with are hard to measure or difficult to control.

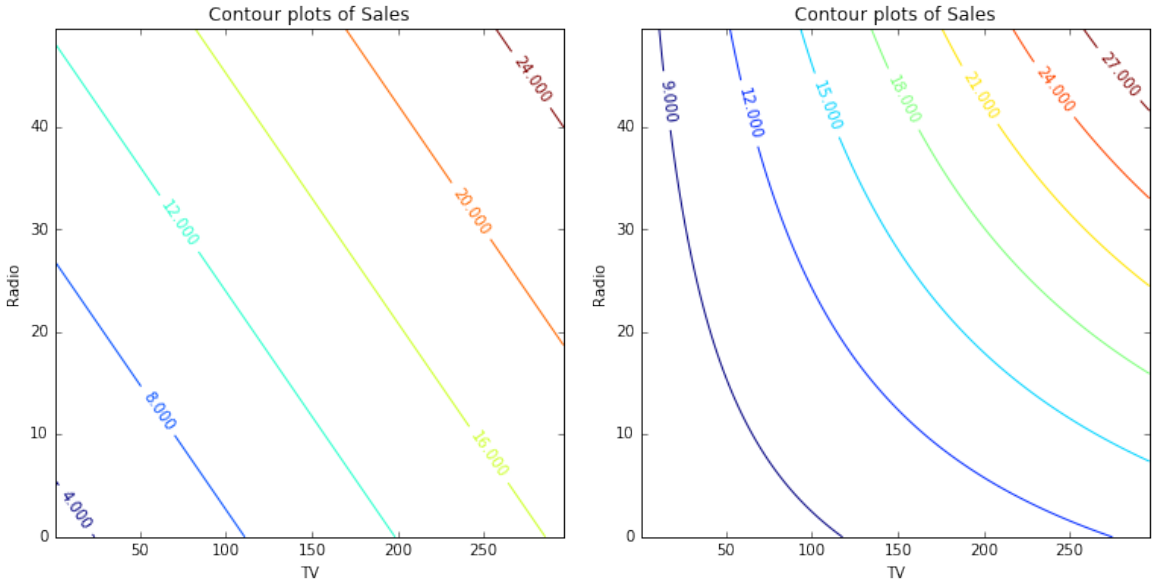
Real-world examples of interaction include:

- Interaction between adding sugar to coffee and stirring the coffee. Neither of the two individual variables has much effect on sweetness but a combination of the two does.
- Interaction between adding carbon to steel and quenching. Neither of the two individually has much effect on strength but a combination of the two has a dramatic effect.
- Interaction between smoking and inhaling asbestos fibres: Both raise lung carcinoma risk, but exposure to asbestos multiplies the cancer risk in smokers and non-smokers. Here, the joint effect of inhaling asbestos and smoking is higher than the sum of both effects.
- Interaction between genetic risk factors for type 2 diabetes and diet (specifically, a "western" dietary pattern). The western dietary pattern was shown to increase diabetes risk for subjects with a high "genetic risk score", but not for other subjects.

To recognize the possible interactions between two variables, we can explore their relation graphically. There are three major types of interactions.

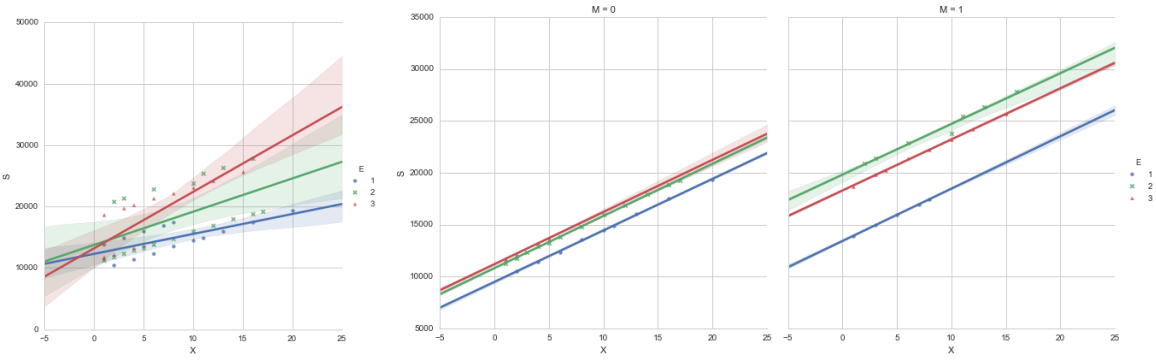
1. Interaction between two continuous variables

   If there is no interaction, the mean response is a plane (linear in both variables). However, when interaction exists, the mean response becomes a curved surface. Alternatively, the contour plots are parallel lines without interaction, and are curves with interactions.



2. Interaction between a continuous variable and an attribute variable

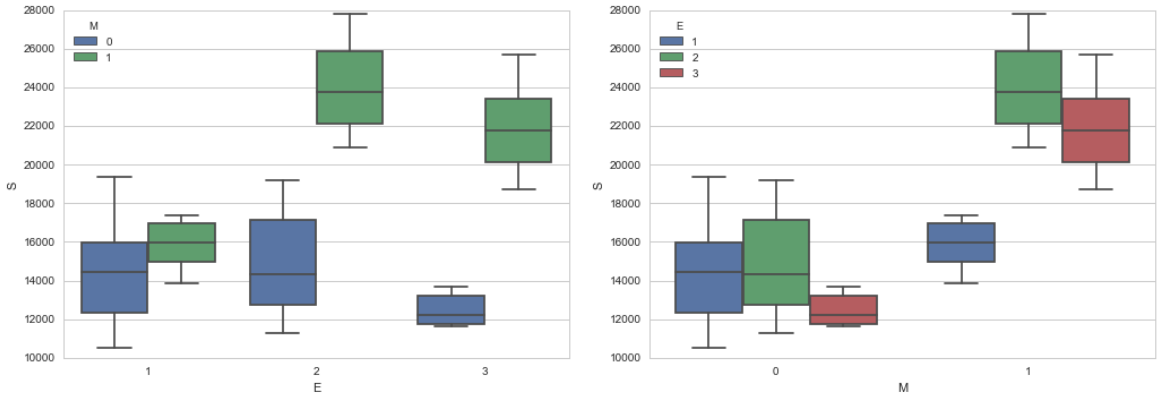
   If there exists interactions between attribute variable and continuous variables, it means the effect of the continuous variable depends on the value of the attribute. In other words, the coefficients (or slope on the graph) are different when the attribute taking different values, as shown below (left figure). On the other hand, if the slopes do not change with the attribute value, there is no significant interaction (the right two figures).



In the left picture, Experience (X) and education level (E) seems to be interactive because the slope of experience depends on the value of education level. However, in the right two figures we can find out that the real interaction exists between education level and management (M). The slope of experience doesn't depends on the value of management or education level.

3. Interaction between two attribute variables

When the effect of one variable depends on the value of another variable, the interaction might exist. Graphically, the pattern of the box plot changes as the other variable takes different value. In both cases below, the pattern of the box plots of the same color changes for different color.



In conclusion, a reasonable sample will be $Y = \beta_0 + \beta_1 \cdot E + \beta_2 \cdot M + \beta_3 \cdot X + \beta_{12} E * M$

```
salaryfit = smf.ols(formula="S~C(E)*C(M)+X", data=salary).fit()
```

	coef	std err	t	P>|t|
Intercept	9472.6854	80.344	117.902	0.000
C(E)[T.2]	1381.6706	77.319	17.870	0.000
C(E)[T.3]	1730.7483	105.334	16.431	0.000
C(M)[T.1]	3981.3769	101.175	39.351	0.000
C(E)[T.2]:C(M)[T.1]	4902.5231	131.359	37.322	0.000
C(E)[T.3]:C(M)[T.1]	3066.0351	149.330	20.532	0.000
X	496.9870	5.566	89.283	0.000

3.3.3 Least square estimation

When n observations are collected $(\mathbf{x}_i, y_i)$, we can estimate the model parameters to identify influential variables or to make forecastings. Following the same criterion to minimize the MSE, we can estimate the parameters by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})^2.$$

The multiple linear regression model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

and the least square criterion reduces to minimizing the vector norm of the difference

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2, \tag{3.15}$$

where $||\mathbf{Y}|| = \sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}$ is the 2-norm of the vector. Using matrix calculus (https://en.wikipedia.org/wiki/Matrix_calculus), we can show that $\hat{\boldsymbol{\beta}}$ again has analytical expression

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{3.16}$$

where $\mathbf{X}^T, \mathbf{X}^{-1}$ represents the transpose and inverse of the matrix, respectively. $\hat{\boldsymbol{\beta}}$ is unique as long as $\mathbf{X}^T\mathbf{X}$ is full rank (or invertible).

Example: In simple linear regression, we can express them in matrix form by

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (n\bar{x})^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}.$$

With some algebraic transformation, we can get consistent results as those in Section 3.2. □

The natural way to estimate the variance $\hat{\sigma}^2$ is

$$\hat{\sigma}^2 = \frac{||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2}{n - p - 1}.$$

Again the term $||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2$ or equivalently $\sum_{i=1}^n (y_i - \mathbf{x}_i \cdot \hat{\boldsymbol{\beta}})^2$ is called sum of square errors (SSE).

In fact, $\hat{\boldsymbol{\beta}}$ is generally a good way to estimate the model parameters as stated in the following theorem.

Theorem 3.1 (Gauss-Markov) *In a linear regression model, in which ϵ_i have expectation zero, equal variances, and are uncorrelated, the ordinary least square estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE). Furthermore, if ϵ_i are normally distributed, $\hat{\boldsymbol{\beta}}$ is the best among* all *unbiased estimators.*

Similar to the case in simple linear regression, when the data are assumed normally distributed, we can get the distribution of $\hat{\boldsymbol{\beta}}$. In more details, $\hat{\boldsymbol{\beta}}$ follows multivariate normal distribution

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2). \tag{3.17}$$

Note that this result not only gives the marginal distribution of each component of $\hat{\boldsymbol{\beta}}$, it also provides the covariance among different components. The results in the simple linear regression are in fact its special case.

There are some remarks about the estimation of $\hat{\boldsymbol{\beta}}$: (1) The accuracy of $\hat{\boldsymbol{\beta}}$ depends on the sample size, and also the scatters of $\mathbf{x}_i$; (2) Sometimes, $\mathbf{X}^T\mathbf{X}$ is singular, then $\boldsymbol{\beta}$ is not estimable. The coefficients are not interpretable (collinearity); (3) Linear transformation of $\hat{\boldsymbol{\beta}}$ is still normal: $\mathbf{A}\hat{\boldsymbol{\beta}} \sim \text{MVN}(\mathbf{A}\boldsymbol{\beta}, \mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\sigma^2)$.

3.3.4 Confidence & prediction intervals

With the estimated $\hat{\boldsymbol{\beta}}$ and its covariance $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$, and the $\hat{\sigma}^2$, we can construct the confidence interval for the parameters. For each component $\hat{\beta}_j$ in $\hat{\boldsymbol{\beta}}$, its $1-\alpha$ confidence interval can be expressed as

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{d_{ii}}, \tag{3.18}$$

where d_{ii} is the (i,i)th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$. In other words, the standard deviation of $\hat{\beta}_j$ is simply $\sigma\sqrt{d_{ii}}$.

More importantly, with the covariance matrix of $\hat{\boldsymbol{\beta}}$, we can construct the confidence interval (or region) for multiple components of $\hat{\boldsymbol{\beta}}$ together, or some linear combinations of the components.

- What is the 95% joint confidence region for (β_0, β_1)?
- What is the confidence interval for the difference $\beta_1 - \beta_2$?

A general way can be found using the property of multivariate normal distribution. For any matrix $\mathbf{A}$ with rank q, we know that

$$\mathbf{A}\hat{\boldsymbol{\beta}} \sim \text{MVN}(\mathbf{A}\boldsymbol{\beta}, \mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T\sigma^2).$$

Consequently, we can further derive

$$\frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})^T[\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{A}\boldsymbol{\beta})}{q\hat{\sigma}^2} \sim F_{q,n-p-1},$$

where $F_{q,n-p-1}$ is the F distribution with degree of freedom q and $n-p-1$. The $1-\alpha$ level confidence region for $\mathbf{A}\boldsymbol{\beta}$ is thus defined as the set of any q dimensional points satisfying

$$\mathbf{A}\boldsymbol{\beta} : \left\{ \mathbf{b} \in \mathbb{R}^q : (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b})^T [\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T]^{-1} (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{b}) \leq q\hat{\sigma}^2 \cdot F_{q,n-p-1,\alpha} \right\}, \tag{3.19}$$

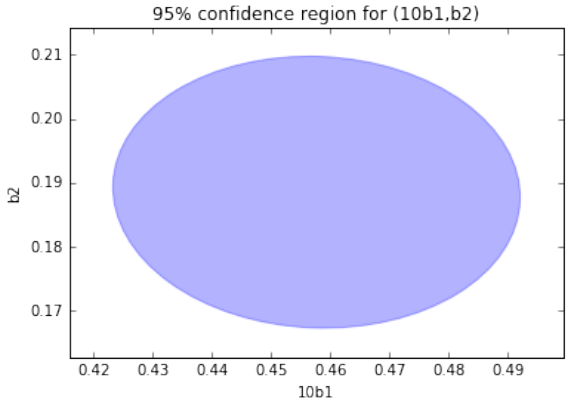
where $F_{q,n-p-1,\alpha}$ is the upper α quantile of the F distribution.

Example: The confidence region (3.19) is very general to include many useful cases as special case. We use the following examples to demonstrate.

I. The confidence region for all the parameters $\boldsymbol{\beta}$. In this case, $\mathbf{A} = \mathbf{I}_{p+1}$ with rank $p+1$, and the confidence region becomes:

$$\boldsymbol{\beta} : \left\{ \mathbf{b} \in \mathbb{R}^q : (\hat{\boldsymbol{\beta}} - \mathbf{b})^T (\mathbf{X}^T\mathbf{X}) (\hat{\boldsymbol{\beta}} - \mathbf{b}) \leq (p+1)\hat{\sigma}^2 \cdot F_{p+1,n-p-1,\alpha} \right\}.$$

When $\hat{\boldsymbol{\beta}}$ has two dimension, this region is an ellipse in the plane.

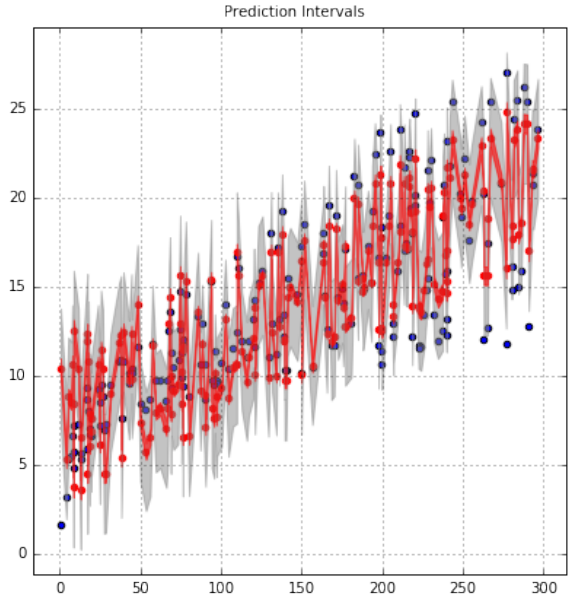


II. The confidence interval for mean response at new predictor value $\mathbf{x}^*$. The point forecast is straightforward, with mean response $y^* = \mathbf{x}^*\hat{\boldsymbol{\beta}}$ based on (3.13). To further obtain the confidence interval for the values, we can use (3.19) with $\mathbf{A} = \mathbf{x}^*$ of rank $q = 1$. Hence, the confidence interval of the mean response satisfy

$$y : \left\{ y \in \mathbb{R} : (y^* - y)^T [\mathbf{x}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*]^{-1} (y^* - y) \leq \hat{\sigma}^2 \cdot F_{1,n-p-1,\alpha} \right\}.$$

Equivalently, using the relation between $F_{1,n-p-1}$ and t_{n-p-1}, we can get a more explicit form

$$y : y^* \pm t_{n-p-1,\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}^*(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}^*}.$$



III. Other useful comparisons. For example, if β_1 and β_2 represents coefficients of two predictor variables. $\beta_1 - \beta_2$ is a measure of their relative effects on the response. To get the confidence region for $(\beta_1 - \beta_2, \beta_2 - \beta_3)$, we can define $\mathbf{A}$ with rank $q = 2$ as

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 \end{bmatrix},$$

and then follow the formula in (3.19).

4 Model Checking and Diagnosis

Major assumptions of linear regression include

1. The relationship between the outcomes and the predictors is (approximately) linear.
2. The error term ϵ has zero mean.
3. The error term ϵ has constant variance.
4. The errors are uncorrelated.
5. The errors are normally distributed or we have an adequate sample size to rely on large sample theory.

We should always check the fitted models to make sure that these assumptions have not been violated.

4.1 Residuals

The diagnostic methods we'll be exploring are based primarily on the residuals. Recall, the residual is defined as

$$e_i = y_i - \hat{y}_i, \quad i = 1, ..., n$$

where $\hat{y}_i = \mathbf{X}\hat{\boldsymbol{\beta}}$. If the model is appropriate, it is reasonable to expect the residuals to exhibit properties that agree with the stated assumptions.

According to the definition of the residuals, it is easy to show that the mean of the residuals is 0,

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0,$$

and it can yield the estimation of the population variance

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2.$$

Precisely speaking, The $e_i, i = 1, \cdots, n$ are not independent random variables. In general, if the number of residuals (n) is large relative to the number of predictor variables (p), the dependency can be ignored for all practical purposes in an analysis of residuals.

To analyze the residuals in different contexts, it is also common to "standardize" the residuals by dividing its standard deviation. Using matrix form, we can write the residual vector as

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}.$$

The term $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is often called hat matrix, and plays a crucial role in linear regression analysis and model diagnostics. Using the notation, we can get the covariance matrix of $\mathbf{e}$ is simply $(\mathbf{I} - \mathbf{H})\sigma^2$. As a result, the **studentized residual** is defined as

$$r_i = \frac{e_i}{\sqrt{1 - h_{ii}}\hat{\sigma}}, \tag{4.1}$$

where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H}$. When the assumptions of the linear model hold, r_i *follows a t distribution with* $n - p - 1$ *degree of freedom.* Consequently, r_i is free from measurement scales in different contexts.

4.2 Diagnostics using residuals

The model assumptions can be checked against the property of the residuals.There are two kinds of residual analysis:

1. Major graphical tools for model checking

- QQ plot to check normality
- Scatter plot to check linearity and variance
- Autocorrelation plot to check independence

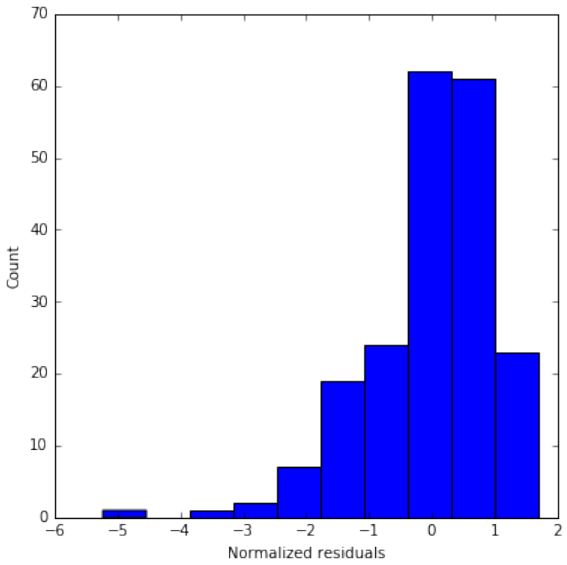
2. Specialized hypothesis testing on residuals

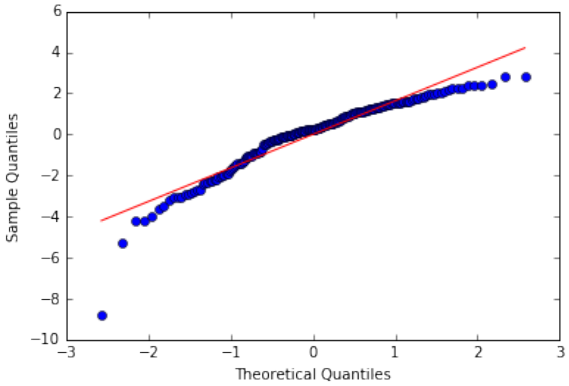
4.2.1 Major graphical tools

Residual analysis is usually done graphically. We describes the major plots as follows.

I. Normal probability plot (or quantile-quantile plot).

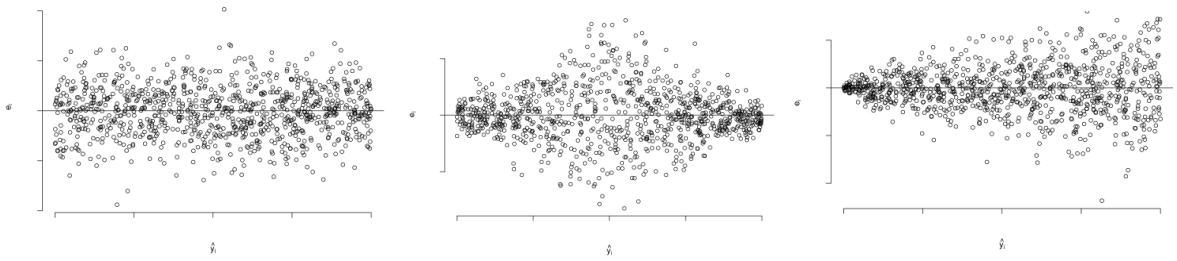
Using quantile-quantile (QQ) plot, we can compare quantiles of a sample to the expected quantiles if the sample came from some distribution for a visual assessment. To construct a quantile-quantile plot for the residuals, we plot the quantiles of the residuals against the theoretical quantiles of a *normal distribution*. If the residuals follow a normal distribution, the QQ plot should resemble a straight line. A straight line connecting the 1st and 3rd quartiles is often added to the plot to aid in visual assessment.





II. Scatter plots

Another useful aid for inspection is a scatter plot of the residuals against the fitted values and/or the predictors. These plots can help us identify: non-constant variance, violation of the assumption of linearity, and potential outliers.



Non-constant variance can often be remedied using appropriate transformations. Ideally, we would choose the transformation based on some prior scientific knowledge, but this might not always be available. Some typical choices are listed below

Relation of σ^2 to $\mathbb{E}(Y\|\mathbf{x})$	Transformation	Comment
$\sigma^2 \propto$ constant	$y' = y$	no transformation
$\sigma^2 \propto \mathbb{E}(Y)$	$y' = \sqrt{y}$	Poisson data
$\sigma^2 \propto \mathbb{E}(Y)^2$	$y' = \ln(y)$	$y > 0$

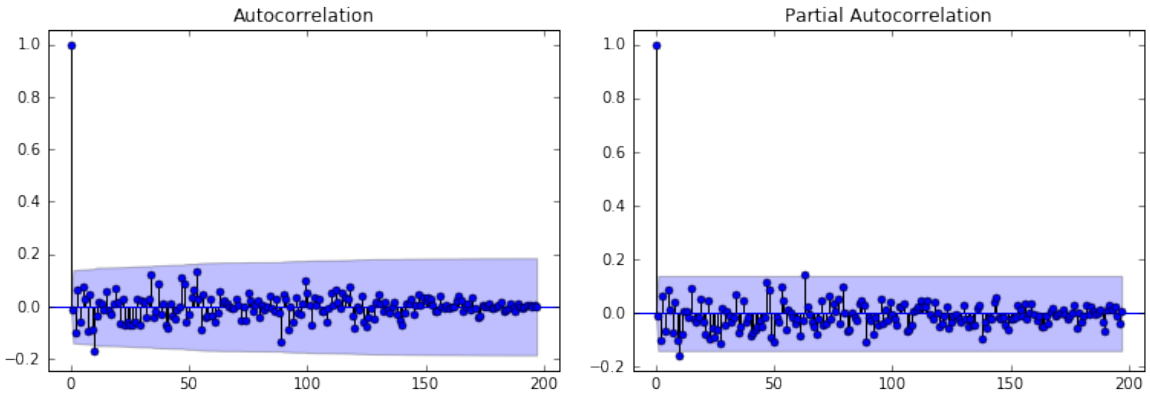
In general, for $Y > 0$, an automatic transformation can be done (suggested) by Box-Cox transformation, in the form

$$y' = \begin{cases} (y^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \ln(y), & \lambda = 0 \end{cases} . \tag{4.2}$$

The best choice of λ can be determined based on the data.

III. Independence check/test

If the samples are independent, the residuals should not have visible patterns when plotted against time or observation index. Autocorrelation plot or partial autocorrelation plot (will be discussed later) can graphically illustrate the degree of violation. Some formal statistical tests have also been developed to test the independence of the data.



4.2.2 Tests for certain property

I. Shapiro-Wilk test for normality

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2},$$

where $e_{(i)}$ is the ith smallest value, a_i are some constants.

II. Modified Levene Test for constant variance

III. Durbin-Watson statistic In statistics, the Durbin-Watson statistic is a test statistic used to detect the presence of autocorrelation in the residuals from a regression analysis. If e_t is the residual associated with the observation at time t, then the test statistic is

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

where T is the number of observations. Since d is approximately equal to $2(1 - r)$, where r is the sample autocorrelation of the residuals, $d = 2$ indicates no autocorrelation. The value of d always lies between 0 and 4. If the Durbin-Watson statistic is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin-Watson is less than 1.0, there may be cause for alarm.

4.3 Outliers, leverage points, influential points, collinearity

4.3.1 Identifying outliers

An outlier is an extreme observation. Depending on their location in the predictor space, outliers can have severe effects on the regression model. We can use jackknife residuals to identify potential outliers. Any points that are greater than 3 or 4 standard deviations away from 0 may be considered potential outliers.

There are several scenarios for outliers.

- "Bad" data that results from unusual but explainable events, eg - malfunction of measuring instrument, incorrect recording of data. In this case we should try to retrieve the correct value, but if that's not possible we may need to discard the data point.
- Inadequacies in the model. The model may fail to fit the data well for certain values of the predictor. In this case it could be disastrous to simply discard outliers.
- Poor sampling of observations in the tail of the distribution. This may be especially true if the outcome arises from a heavy-tailed distribution.

With a sample size of 60, we might expect 2 or 3 residuals to be further than 2 standard deviation from 0 and none to be more than 3 standard deviation.

4.3.2 High leverage points

Leverage is a measure of how strongly the data for obs i determine the fitted value $\hat{Y}_i$. If h_{ii} is close to 1, the fitted line will usually pass close to $(\mathbf{x}_i, Y_i)$.

The hat matrix,

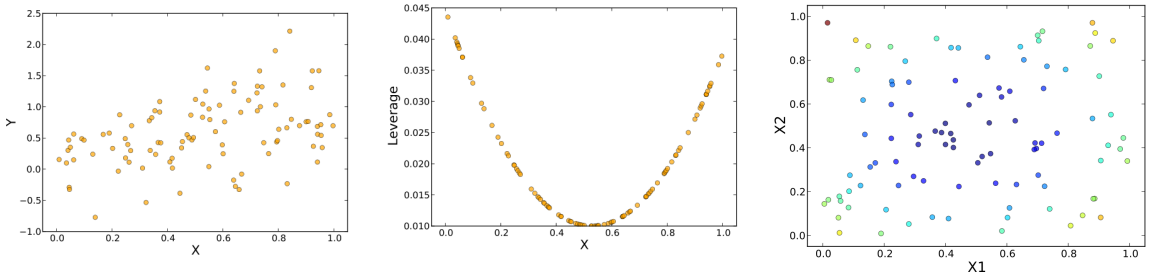
$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

plays an important role in identifying influential observations. The diagonal elements $h_{ii} = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}x_i^T$, where $\mathbf{x}_i$ is the ith row of the $\mathbf{X}$ matrix, play an especially important role. h_{ii} is a standardized measure of the distance of the covariate values for ith observation and the means of the $\mathbf{X}$ values for all n observations.

Also,

$$0 \leq h_{ii} \leq 1, \quad \sum_{i=1}^{n} h_{ii} = p + 1.$$

Therefore the average size of a hat diagonal is $\bar{h} = (p+1)/n$. Leverage values greater than $2\bar{h}$ are considered to be high leverage with regard to their $\mathbf{x}_i$ values and we would consider them high leverage points. The left two pictures below shows the leverage values in a simple linear regression. The third picture shows leverage values in a multiple linear regression.



4.3.3 Identifying influential observations

Points that are remote in the predictor space may not influence the estimate of the regression coefficients but may influence other summary statistics, such as R^2 and the standard errors of the coefficients. These points are called **leverage points**. Points that have a noticeable effect on the regression coefficients are called **influential points**. In other words, **Influence** measures the degree to which deletion of an observations changes the fitted model. A high leverage point has the potential to be influential, but is not always influential

Influence can be measured by Cook's distance. Cook's Distance measures the influence of the ith observation on all n fitted values and is given by

$$D_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(-i)})^T(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(-i)})}{(p+1)\hat{\sigma}^2}$$

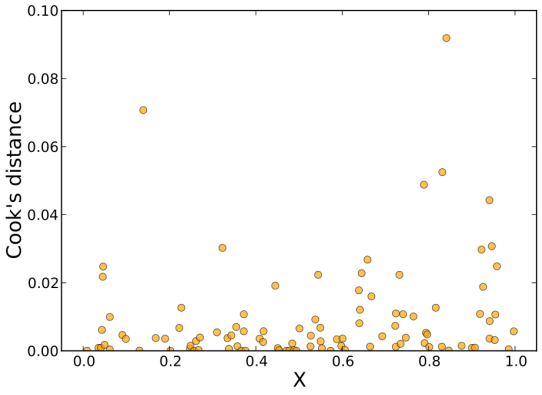
where $\hat{\mathbf{Y}}$ is the vector of fitted values when all n observations are included and $\hat{\mathbf{Y}}_{(-i)}$ is the vector of fitted values when the ith observation is deleted. Cook's D can also be expressed as

$$D_i = \frac{e_i^2}{(p+1)\hat{\sigma}^2} \frac{h_{ii}}{(1-h_{ii})^2}$$

From this expression we see that D_i depends on both the size of the residual e_i, and the leverage, h_{ii}.

The magnitude of D_i is usually assessed by comparing it to $F_{p+1,n-p-1}$. If the percentile value is less than 10 or 20 %, then the ith observation has little apparent influence on the fitted values. If the percentile value is greater than 50%, we conclude that the ith observation has significant effect on the fitted values.

As a general rule, D_i values from 0.5 to 1 are high, and values greater than 1 are considered to be a possible problem.



4.3.4 Collinearity

In statistics, multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy. In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

Indicators that multicollinearity may be present in a model:

- Large changes in the estimated regression coefficients when a predictor variable is added or deleted
- Insignificant regression coefficients for the affected variables in the multiple regression, but a rejection of the joint hypothesis that those coefficients are all zero (using an F-test)

- If a multiple regression finds an insignificant coefficient of a particular explanatory variable, yet a simple linear regression of the explained variable on this explanatory variable shows its coefficient to be significantly different from zero, this situation indicates multicollinearity in the multiple regression
- Some authors have suggested a formal variance inflation factor (VIF) for multicollinearity:

$$\text{VIF} = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination of a regression of explanatory variable j on all the other explanatory variables:

$$X_j = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \alpha_p X_p + \varepsilon,$$

The better the fit, the more severe the collinearity. A VIF of 5 or 10 and above indicates a multicollinearity problem.

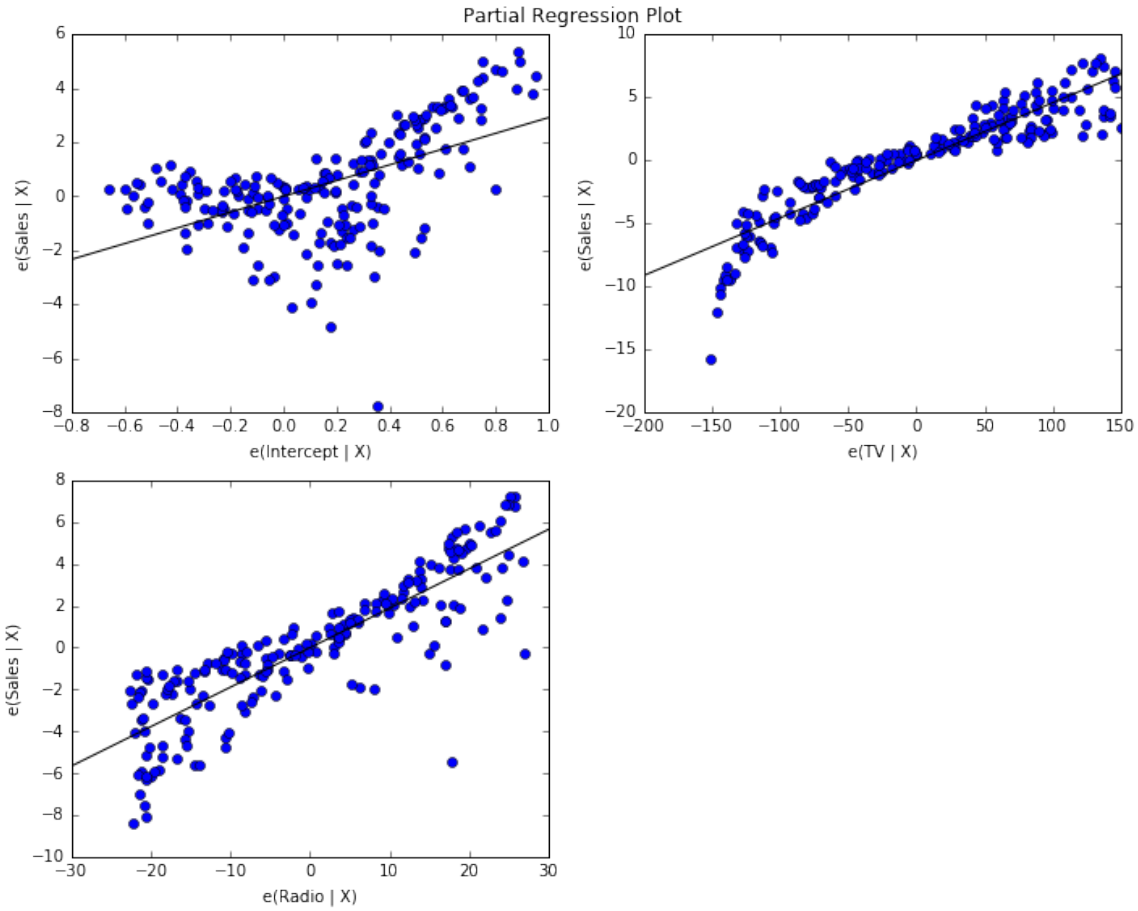
Added variable plots are also called **partial regression plots**, or adjusted variable plot. It allows us to study the marginal relationship of a regression given the other variables that are in the model. For the variable X_j

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \beta_p X_p + \epsilon$$

and

$$X_j = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_{j-1} X_{j-1} + \alpha_{j+1} X_{j+1} + \alpha_p X_p + \varepsilon$$

Plot $Y - \hat{Y}$ vs $X_j - \hat{X}_j$



Some comments on using the plots

- They only suggest possible relationships between the predictor and the response.
- In general, they will not detect interactions between regressors.
- The presence of strong multicollinearity can cause partial regression plots to give incorrect information

5 Model Evaluation and Selection

Regression models have two major objectives: i) quantifying the effects of each predictors, considering the influences of other predictors; ii) predict the (mean) response at other unobserved predictor values. It is important to evaluate the regression models, and select the "best model" among all candidates to make the forecasting more accurate. Here are a few reasons why we want to select the "best model".

1. We want to explain the data in the simplest way. Redundant predictors should be removed. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that the smallest model that fits the data is best.
2. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Degrees of freedom will be wasted.
3. Collinearity is caused by having too many variables trying to do the same job.
4. Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

5.1 Evaluating the regression model

In the introduction, we discussed the criteria to evaluate the forecasting performance, including MSE and MAD. In the regression context, there are several evaluation criteria developed based on them. We will go through them as follows.

I **The "notorious"** R^2:

R^2, also called *coefficient of determination*, evaluates the percentage of total variation (uncertainty) explained by the regression model. Mathematically, it is defined as:

$$R^2 = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{5.1}$$

Compared with MSE definition, we can see that MSE = SSE/n. In other words, the smaller the MSE, the closer R^2 to 1. It appears R^2 is a good measure of the forecasting performance. Unfortunately, there is an inherent problem: the forecasting errors are calculated on the same dataset as that used for model estimation. As a result, it often under-estimates the forecasting errors when used in future predictions. In fact, by increasing the number of predictors (relevant or not), R^2 always increases. As a result, it is *never* used as a criterion to select the "best" model because only the largest model has the largest R^2.

II **Adjusted** R^2

Since R^2 always increases as the model size increases, an adjusted R^2 is proposed, often denoted by R_a^2. It is defined by

$$R_a^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{TSS}/(n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}. \tag{5.2}$$

Because of the adjustment, increasing the model size, will increase R^2, but not necessarily increase R_a^2. Adding a predictor will only increase R_a^2 if it has some value in prediction. From

another angle, minimizing the standard error for prediction means maximizing $R_2^2 a$. Compared with R^2, it "penalized" bigger models.

III **Cross-validated forecast errors**:

While the MSE and MAD are intuitive criteria to evaluate the forecasting performance, the difficulty lies in how to obtain them accurately. Using the same dataset to estimate the model and to calculate the MSE (e.g., R^2) cannot provide reliable assessment. A natural solution is to use two independent dataset, one *training* set for model estimation, and the other *testing* set for model evaluation.

Cross validation is one such strategy to evaluate the forecasting errors more reliably. For k-fold cross validation, it often consists of the following steps

(a) Randomly divide the data into k non-overlapping subsets, of (roughly) equal size.

(b) Select one subset as the testing data, and the remaining $k-1$ subsets combined as training data. Estimate the model using the training data, and compute the prediction error (e.g., MSE) on the testing data, denoted by MSE_i.

(c) Repeat this procedure for k times, with each of the k subsets being the testing data.

(d) Average the k prediction error estimates to get the cross-validated error $\text{MSE}_{CV} = \sum_{j=1}^{k} \text{MSE}_j/k$.

Compared with other criteria, cross validated forecast error is more intuitive and often more effective. However, it requires much more computational effort (k times), except certain special cases. Common choice of k includes $k = 5, k = 10$. When $k = n$, it is more commonly known as *leave-one-out cross validation*.

IV **Akaike's Information Criterion (AIC), Schwarz's BIC**

The information criteria, including Akaike's Information Criterion (AIC) and Schwarz's Bayesian Information Criterion (BIC), are commonly used for model comparison or selection. For linear regression models, they can be reduced to

$$\text{AIC} = n \ln\left(\frac{\text{SSE}}{n}\right) + 2(p+1) \tag{5.3}$$

$$\text{BIC} = n \ln\left(\frac{\text{SSE}}{n}\right) + (p+1)\ln(n). \tag{5.4}$$

We want to minimize AIC or BIC to select the "best" model. Larger models will fit better and so have smaller SSE. But they also use more parameters. Thus the "best" model will balance the goodness-of-fit with model size. BIC penalizes larger models more heavily and so tends to prefer smaller models comparing with AIC. AIC and BIC can be used as selection criteria for other types of model (not limited to regression models).

V **Mallow's C_p**

The criterion is developed based on the intuition that: a good model should predict well, so average MSE of the prediction might be small

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \mathbb{E}(\hat{y}_i - \mathbb{E}y_i)^2.$$

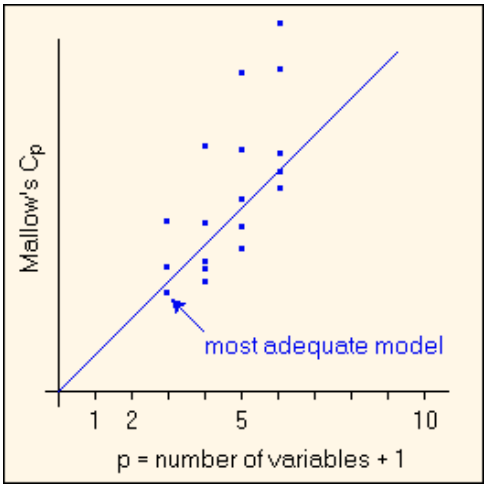
This quantity can be estimated by the C_p statistic

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} + 2(P+1) - n, \tag{5.5}$$

where $\hat{\sigma}^2$ is from the model with all P predictors and SSE_p indicates the sum of squared errors from a model with p *parameters*. In a sense, C_p balances the model errors (in terms of SSE) and the number of predictors used (in terms of p). C_p has the following properties in model selection:

- C_p is easy to compute
- It is closely related to R_a^2 and the AIC.
- For the full model $C_p = P + 1$ exactly.
- If a model with p parameter fits the data, then $\mathbb{E}(C_p) \approx p$. A model with a bad fit will have C_p much larger than p.

It is usual to plot C_p against p. We desire models with small p and C_p around or less than p.



5.2 Selecting the regression model

When we have many predictors (with many possible interactions), it can be difficult to find a good model. It can be challenging to find which main effects do we include, and which interactions do we include. Model selection tries to "simplify" this task. However, this is still an "unsolved" problem in statistics. There are no magic procedures to get you the "best model."

I **All subset selection**

When the number of predictors is not large, it is possible to enumerate all possible models with different number and different set of predictors. When there are m candidate predictors, the total number of distinct model is 2^m, without considering interactions and transformation. Given any criteria (adjusted R^2, Mallow's C_p, BIC, AIC, or cross validation error), we can find the model with optimal value. For models with close performance, the smaller model is preferred.

II **Greedy search**

When the number of predictors becomes large, it is not feasible to conduct *all subset selection*, especially when interactions and transformations of variables should be considered. In this case, some greedy search (or other heuristic methods) should be used to find the "best" model by certain evaluation criterion.

(a) Backward Elimination

This is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run while taking account of what variables are eligible for removal.

- Start with all the predictors in the model;

- Remove the predictor leading to largest improvement in performance;
- Refit the model and goto Step 2;
- Stop when no more improvement can be made by removing predictors;

(b) Forward Selection

Forward selection reverses the backward method.

- Start with no variables in the model;
- For all predictors not in the model, check the model performance if they are added to the model;
- Choose the one leading to largest improvement, and include it in the model;
- Continue until no new predictors can be added.

(c) Stepwise Regression

This is a combination of backward elimination and forward selection. This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done.

Greedy procedures are relatively cheap computationally but they do have some drawbacks.

- Because of the "one-at-a-time" nature of adding/dropping variables, it's possible to miss the "best" model.
- The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest. With any variable selection method, it is important to keep in mind that model selection cannot be divorced from the underlying purpose of the investigation. Variable selection tends to amplify the statistical significance of the variables that stay in the model. Variables that are dropped can still be correlated with the response. It would be wrong to say these variables are unrelated to the response, it's just that they provide no additional explanatory effect beyond those variables already included in the model.
- Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes. To give a simple example, consider the simple regression with just one predictor variable. Suppose that the slope for this predictor is not quite statistically significant. We might not have enough evidence to say that it is related to y but it still might be better to use it for predictive purposes.

6 Hypothesis Testing in Regression Models

A statistical hypothesis is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. A statistical hypothesis test is a method of statistical

inference.

Hypothesis testing can be used to formally test whether a predictor (or its transformation and interaction with other predictors) is statistically significant in predicting the mean response. In the general framework, it includes a few commonly used special cases. Some of the results have been summarized before in different places.

Hypothesis testing allows us to carry out inferences about population parameters using data from a sample. In order to test a hypothesis in statistics, we must perform the following steps: 1) Formulate a null hypothesis and an alternative hypothesis on population parameters; 2) Build a statistic to test the hypothesis made; 3) Define a decision rule to reject or not to reject the null hypothesis.

It is very important to remark that hypothesis testing is always about population parameters. Hypothesis testing implies making a decision, on the basis of sample data, on whether to reject that certain restrictions are satisfied by the basic assumed model. The restrictions we are going to test are known as the null hypothesis, denoted by H_0. Thus, null hypothesis is a statement on population parameters.

The details of testing process shows below.

1. State the relevant null and alternative hypotheses.
2. Consider the statistical assumptions being made about the sample in doing the test;
3. Decide and state the relevant test statistic T.
4. Derive the distribution of the test statistic under the null hypothesis
5. Select a significance level α.
6. Compute from the observations the observed value of the test statistic T.
7. Decide to either reject the null hypothesis in favor of the alternative or not reject it.

I **Testing a single** $\beta_j = 0$

Using the results in multiple linear regression in Chapter 3, we know that the least square estimation $\hat{\beta}_j$ follows normal distribution with mean β_j and corresponding variance. In addition, we have

$$\frac{\hat{\beta}_j - \beta_j}{\text{SE}(\hat{\beta}_j)} \sim t_{n-p-1},$$

when there are p predictors in the model. Therefore, the natural way to test $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is to use the statistic $T = \hat{\beta}_j/\text{SE}(\hat{\beta}_j)$, with decision rule

$$|T| : \begin{cases} > t_{n-p-1,\alpha/2}, & \text{Reject null hypothesis} \\ \leq t_{n-p-1,\alpha/2}, & \text{Do not reject null hypothesis} \end{cases},$$

where α is the significance level. For most regression outputs, the test values besides each predictor indicate such test results

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1.5925	1.389	1.146	0.252	-1.130	4.315
x1	-0.0016	0.001	-1.072	0.284	-0.004	0.001
x2	0.0006	0.008	0.073	0.942	-0.015	0.016
x3	9.349e-05	4.02e-05	2.323	0.020	1.46e-05	0.000
x4	-0.0003	0.008	-0.035	0.972	-0.016	0.016
x5	-0.0001	4.51e-05	-2.694	0.007	-0.000	-3.31e-05
x6	-0.0001	9.08e-05	-1.184	0.236	-0.000	7.04e-05
x7	6.249e-08	5.81e-08	1.076	0.282	-5.14e-08	1.76e-07
x8	1.96e-06	2.53e-06	0.774	0.439	-3e-06	6.92e-06
x9	-0.0006	0.000	-2.958	0.003	-0.001	-0.000

It is also noted that the test significance shall be interpreted one by one. Because the test is equivalent to the following test

$$H_0 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{j-1} x_{j-1} + \qquad \beta_{j+1} x_{j+1} + \cdots + \beta_p x_p + \epsilon$$
$$H_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \cdots + \beta_p x_p + \epsilon$$

II **Testing all predictors simultaneously**

Testing model significance, or overall significance, is a particular case. Model significance means global significance of the model. One could think that the test is formulated in the following

$$H_0 : Y = \beta_0 + \epsilon$$
$$H_1 : Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \cdots + \beta_p x_p + \epsilon.$$

In other words, none of the predictors need to be included in the model. We can use the sum of square errors (SSE), or equivalently R^2 to express the test statistic

$$T = \frac{R^2/p}{(1 - R^2)/(n - p - 1)} \sim F_{p,n-p-1}.$$

As a result, if $T > F_{p,n-p-1,\alpha}$, we should reject this hypothesis. Usually, the test result is also

prepared in the software output.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.000
Method:	Least Squares	F-statistic:	2.676
Date:	Mon, 12 Sep 2016	Prob (F-statistic):	0.00417
Time:	22:12:18	Log-Likelihood:	-28217.
No. Observations:	41407	AIC:	5.645e+04
Df Residuals:	41397	BIC:	5.654e+04
Df Model:	9		
Covariance Type:	nonrobust		

III **Testing a sub-model nested in the full model**

More often than not, we are interested in testing some hypotheses lying in between the above two cases. They are identical in mathematical formulation, but might imply different computational cost. The following are some types of such tests on regression parameters.

$$
\begin{aligned}
H_0 :&\beta_2 = \beta_5 = \beta_p = 0 \\
H_0 :&\beta_1 + \beta_2 = 1 \\
H_0 :&\beta_1 = 0 \\
&\beta_2 = 1 \\
&\beta_{p-1} + \beta_p = 0
\end{aligned}
$$

Note that the first two cases only involve one constraint on the regression parameters. The third example has 3 constraints on the regression models. For this group of tests, we can still use F test with corresponding degree of freedom.

In more details, we can express the null hypothesis in a general way as $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$, with row rank of $\mathbf{A}$ being m. Then we can solve the constrained least square estimation

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2, \quad s.t. \;\; \mathbf{A}\boldsymbol{\beta} = \mathbf{c} \tag{6.1}$$

If the null hypothesis is correct, we would expect that the constrained model and unconstrained model has similar performance in terms of estimation of $\boldsymbol{\beta}$ or the approximation difference

$\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$. As a result, the F test can be constructed as

$$T = \frac{(||\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}||^2 - ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2)/m}{||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2/(n-p-1)} \sim F_{m,n-p-1}.$$

If $T > F_{m,n-p-1,\alpha}$, we reject the null hypothesis, meaning that the constrained model is not sufficient in explaining the variation of the response. This test can be done by fitting two models and use ANOVA to get the test results.

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	765.0	13038.806074	0.0	NaN	NaN	NaN
1	762.0	6617.423890	3.0	6421.382184	246.47523	9.268399e-112

Example of meaningful linear constraints on parameters

To examine whether there are constant returns to scale in the chemical sector, we are going to use the Cobb-Douglas production function, given by

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + \epsilon$$

In the above model parameters β_2 and β_3 are elasticities (output/labor and output/capital).

Before making inferences, remember that returns to scale refers to a technical property of the production function examining changes in output subsequent to a change of the same proportion in all inputs, which are labor and capital in this case. If output increases by that same proportional change then there are constant returns to scale. Constant returns to scale imply that if the factors labor and capital increase at a certain rate (say 10%), output will increase at the same rate (e.g., 10%). If output increases by more than that proportion, there are increasing returns to scale. If output increases by less than that proportional change, there are decreasing returns to scale. In the above model, the following occurs

- If $\beta_2 + \beta_3 = 1$, there are *constant* returns to scale
- If $\beta_2 + \beta_3 > 1$, there are *increasing* returns to scale
- If $\beta_2 + \beta_3 < 1$, there are *decreasing* returns to scale.

To answer the question posed in this example, we must test

$$H_0: \ \beta_2 + \beta_3 = 1, \quad v.s. \quad H_1: \ \beta_2 + \beta_3 \neq 1.$$

Reference

In economics, elasticity is the measurement of how responsive an economic variable is to a change in another. Elasticity can be quantified as the ratio of the percentage change in one variable to

the percentage change in another variable, when the latter variable has a causal influence on the former. The elasticity on response on predictor x_i can be calculated easily from the marginal effect

$$e_i \equiv \frac{d(\ln Y)}{d(\ln x_i)}.$$

7 Methods Beyond Linear Regression

Beyond classical regression models, there are also some extensions developed. An incomplete list is provided below:

- Generalized least square method when error variance are not constant, also refered as regression with heterogenous variance.
- Robust regression is designed for non-normal ϵ with unknown distributions. They are robust to outliers or extreme values.
- Nonparametric regression is used to model unknown relation between covariates and responses, which goes beyond linear assumptions.
- Quantile regression investigates the relation between covariates and quantiles of response (not the mean of the response). It has wide application in economics and social science studies.

In addition to regression, there are many methods developed to classify an entity into certain category based on its covariate values.

Definition 7.1 *In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.*

It has widespread application in different domains, and becomes especially hot in current AI buzz. Some typical application includes medical diagnosis, precision medicine; spam email filtering; face recognition; virtual reality, augmented reality; handwriting recognition, voice recognition; recommendation, job screening. In other module, we are *not* going to cover them in details. Instead, we just provide some keywords and examples, and point you to further reference should you need them in the future.

- Logistic regression
- Fisher's linear discriminant analysis (LDA)
- Naive Bayes classifier
- Support vector machines
- k-nearest neighbor

- Boosting
- Decision trees, random forests
- (Deep) neural networks

8 Regression on Time

8.1 Time Series Regression

The dependent variable y is a function time t. It can be modeled as a trend model

$$y_t = TR_t + \epsilon_t,$$

where y_t is the value of the time series in period t; TR_t is the trend in period t, and ϵ_t is the error term in period t. Compared with cross-sectional data, there is no other covariates except time in the model. Depending on the complexity of the trend, it can use

- No trend: $TR_t = \beta_0$
- Linear trend: $TR_t = \beta_0 + \beta_1 t$
- Quadratic trend: $TR_t = \beta_0 + \beta_1 t + \beta_2 t^2$

Aside from the conceptual differences, the model estimation, and prediction methods are the same as in multiple linear regression.

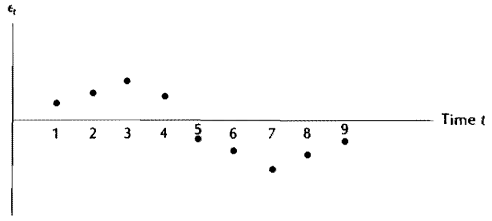
8.2 Detecting Autocorrelation

Compared with cross-sectional data, it is more likely to experience autocorrelation in time series regression. By model assumption, ϵ_t need to be independently and identically distributed. However, we may need to test the validity of the assumption.

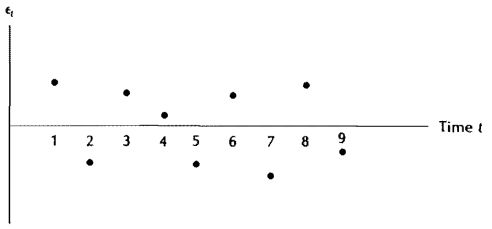
If the relationship between ϵ_t and ϵ_{t-1} can be modeled as

$$\epsilon_t = \phi \epsilon_{t-1} + a_t,$$

it is called first-order autocorrelation when a_t are $i.i.d$ In particular, if $\phi > 0$, ϵ_t has positive correlation; if $\phi < 0$, ϵ_t has negative correlation, and if $\phi = 0$, there is no correlation. We can use residual plot and other diagnostic plot to check



(a) Positive autocorrelation in the error terms: Cyclical pattern



(b) Negative autocorrelation in the error terms: Alternating pattern

Residual plot is intuitive, but subjective. Durbin-Watson Test is one of the rigorous statistical tests can help identify autocorrelations in the residuals. The test is defined as

$$d = \frac{\sum_{i=2}^{n}(e_t - e_{t-1})^2}{\sum_{i=1}^{n} e_t^2}$$

If e_t are positively correlated, d is small; if e_t are negatively correlated, d is large; If there is no correlation, d is in the middle. In particular, cut offs for different hypothesis testings are provided.

(I) H_0: The error terms are not autocorrelated

H_1: The error terms are **positively** autocorrelated

- If $d < d_{L,\alpha}$, reject H_0
- If $d > d_{U,\alpha}$, do not reject H_0
- If $d_{L,\alpha} \leq d \leq d_{U,\alpha}$, inconclusive

(II) H_0: The error terms are not autocorrelated

H_1: The error terms are **negatively** autocorrelated

- If $4 - d < d_{L,\alpha}$, reject H_0
- If $4 - d > d_{U,\alpha}$, do not reject H_0
- If $d_{L,\alpha} \leq (4 - d) \leq d_{U,\alpha}$, inconclusive

(III) H_0: The error terms are not autocorrelated

H_1: The error terms are **autocorrelated**

- If $d < d_{L,\alpha/2}$ or $4 - d < d_{L,\alpha/2}$, reject H_0
- If $d > d_{U,\alpha/2}$ and $4 - d > d_{U,\alpha/2}$, do not reject H_0
- Otherwise, inconclusive

8.3 Seasonal Variation

Time series data is data collected at regular intervals. When there are patterns that repeat over known, fixed periods of time within the data set, such patterns are known as seasonality, seasonal variation, periodic variation, or periodic fluctuations. This variation can be either regular or semi-regular.

Seasonality may be caused by various factors, such as weather, vacation, and holidays and usually consists of periodic, repetitive, and generally regular and predictable patterns in the levels of a time series. Seasonality can repeat on a weekly, monthly or quarterly basis, these periods of time are structured and occur in a length of time less than a year.

Generally, there are **constant (additive) seasonal variation**, where the magnitude of seasonal swing does not depend on the level of the time series. In contrast, for **multiplicative seasonal variation**, the magnitude of seasonal swing is proportional to the average level determined by the trend. When a time series displays multiplicative seasonal variation, we may apply a transformation to the data to produce a transformed series that displays constant seasonal variation.

For a time series that exhibits constant variation, we can use a model of the following form

$$y_t = TR_t + SN_t + \epsilon_t,$$

where the seasonal factor SN_t can be expressed by using dummy variables:

$$SN_t = \beta_{s_1} x_{s_1,t} + \beta_{s_2} x_{s_2,t} + \cdots + \beta_{s_{L-1}} x_{s_{L-1},t},$$

and L is the period of the season.

$$x_{s_1,t} = \begin{cases} 1, & \text{time period t is season 1} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{s_2,t} = \begin{cases} 1, & \text{time period t is season 2} \\ 0, & \text{otherwise} \end{cases}$$

$$x_{s_{L-1},t} = \begin{cases} 1, & \text{time period t is season L-1} \\ 0, & \text{otherwise} \end{cases}$$

Another way to model seasonal trend is to use the trigonometric functions

$$SN_t = \beta_2 \sin\left(\frac{2\pi}{L}t\right) + \beta_3 \cos\left(\frac{2\pi}{L}t\right)$$

or with more frequencies

$$SN_t = \beta_2 \sin\left(\frac{2\pi}{L}t\right) + \beta_3 \cos\left(\frac{2\pi}{L}t\right) + \beta_4 \sin\left(\frac{4\pi}{L}t\right) + \beta_5 \cos\left(\frac{4\pi}{L}t\right)$$

8.4 Growth Curve Models

The regression models we use to describe the trend and seasonal effects are function of time that are linear in the parameters. There are useful models that are not linear in the parameters.

Growth curve model is an example of this case. It is used for long-term or technological forecasting. There are several type of growth curve models are available to model different kinds of time series data.

Consider a growth curve model

$$y_t = \beta_0 \cdot \beta_1^t \cdot \epsilon_t$$

This model is not linear in the parameters. However, with proper transformation,

$$x_t = \ln y_t = \ln(\beta_0) + \ln(\beta_1) \cdot t + \ln(\epsilon_t)$$

we can get an linear regression form.

Based on the model,

$$y_t = \beta_0 \cdot \beta_1^t \cdot \epsilon_t = \beta_1[\beta_0 \cdot \beta_1^{t-1} \cdot]\epsilon_t \approx \beta_1 y_{t-1}\epsilon_t$$

- β_1 indicates the growth rate of the response
- Equivalent form: $y_t = \beta_0 \cdot \exp(\beta_1 \cdot t) \cdot \epsilon_t$
- Typical solution to some differential equations governing underlying dynamics

Some Other useful growth models includes

- Modified exponential curve
$$y_t = s + \alpha e^{\beta t}$$
- Gompertz curve
$$y_t = s \exp(\alpha e^{\beta t})$$

- Logistic curve

$$y_t = \frac{s}{1 + \alpha e^{ct}}$$

9 Exponential Smoothing

In time series regression, the functions are have constant parameters, i.e.,

- TR_t, SN_t etc are *fixed* functions of t
- The variance of ϵ_t does not change over t

This assumption might be valid in short time span, but is questionable in the long run. We might need to update the model (parameters) to account for *unknown* changes. Exponential smoothing is used in such scenario. It weights the observed time series values unequally (also called exponentially weighted moving average (EWMA)). It is most effective when trend (and seasonal factors) of the time series change over time.

9.1 Simple Exponential Smoothing

If the observations follow a constant trend model

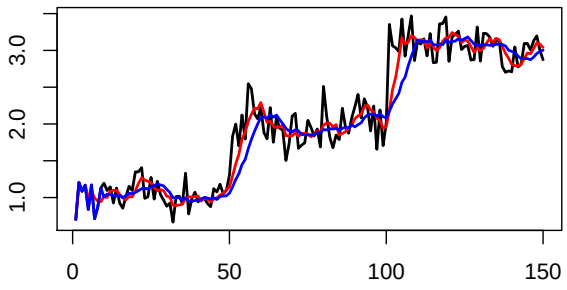
$$Y_t = \beta_0 + \epsilon_t$$

a natural way to estimate β_0 is to take the average

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} Y_i}{n}$$

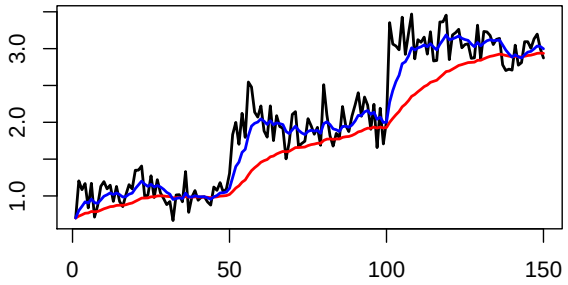
If β_0 is not a constant, but *slowly changing*, then recent observations are more relevant. A simple solution is to take the *moving average*

$$\hat{\beta}_0^n = \frac{\sum_{i=n-w+1}^{n} Y_i}{w}$$



A more popular approach is to use *exponential smoothing*

$$\begin{aligned} L_n &= \alpha Y_n + (1-\alpha) L_{n-1} \\ &= \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} Y_i \end{aligned}$$



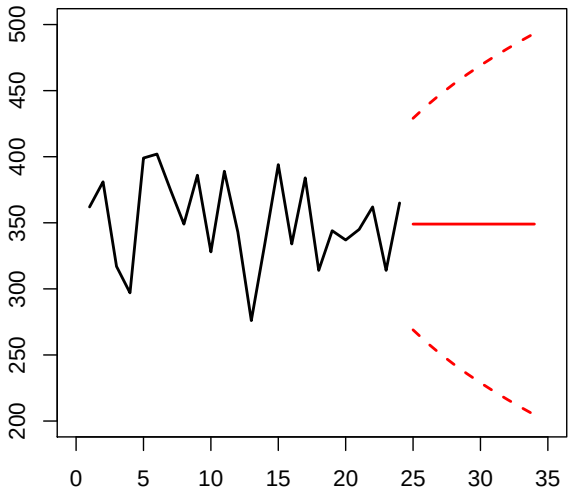
The smoothing constant α is very important. A small α gives smaller weight to current Y_n, leading to smoother curve, slower response to changes. In contrast, a large α gives higher weight to current Y_n, leading to rougher curve, faster response to changes. To select a good α, we can find the value that can minimize the forecast error. Recall that the forecast error at time n can be computed as $e_n = Y_n - L_{n-1}$. Combining the errors together, we have the sum of squared error (SSE):

$$SSE = \sum_{i=1}^{n} (Y_i - L_{i-1})^2$$

Note that SSE depends on α, and as a result, we find the "best" α that can minimize SSE.

Based on the exponential smoothing model, we can forecast future $Y_{n+\tau}$, for $\tau \geq 1$ based on the last information Y_n. Since no trend is assumed, the point forecast equals to L_n. Naturally, the larger τ, the less accurate prediction. We can construct the prediction interval

$$L_n \pm z_{0.025} \cdot s \cdot \sqrt{1 + (\tau - 1)\alpha^2}, \quad s = \sqrt{\frac{SSE}{n-1}}$$



As a summary, we summarized a few common forms of the exponential smoothing.

- Standard form
$$L_n = \alpha Y_n + (1-\alpha)L_{n-1}$$

- Weighted moving average
$$L_n = \alpha Y_n + \alpha(1-\alpha)Y_{n-1} + \cdots + \alpha(1-\alpha)^{n-1}Y_1 + (1-\alpha)^n L_0$$

- Correction form
$$L_n = L_{n-1} + \alpha(Y_n - L_{n-1})$$

9.2 Holt's Trend Corrected Smoothing

The application of simple exponential smoothing is limited, as it allows no trend in the model. Even when linear trend exists in the time series, the simple exponential smoothing might not work. To apply the smoothing in the cases with linear trend, Holt's trend corrected smoothing shall be used.

In more details, consider a time series regression model

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$

If both β_0, β_1 are slowly changing, we need to consider the smoothing for both β_0 and β_1, the intercept and the slope. We can use two smoothings for each parameter, respectively

- Level smoothing
$$L_n = \alpha Y_n + (1-\alpha)(L_{n-1} + B_{n-1}),$$
where B_{n-1} is the estimate of β_1 at step $n-1$

- Growth rate smoothing

$$B_n = \gamma(L_n - L_{n-1}) + (1-\gamma)B_{n-1}$$

The rationale for the second smoothing is due to the observation that from n to $n+1$, the increment in trend function is in fact

$$L_{n+1} - L_n = \beta_0 + \beta_1(n+1) - [\beta_0 + \beta_1 n] = \beta_1,$$

which provides new information on the rate of the trend.

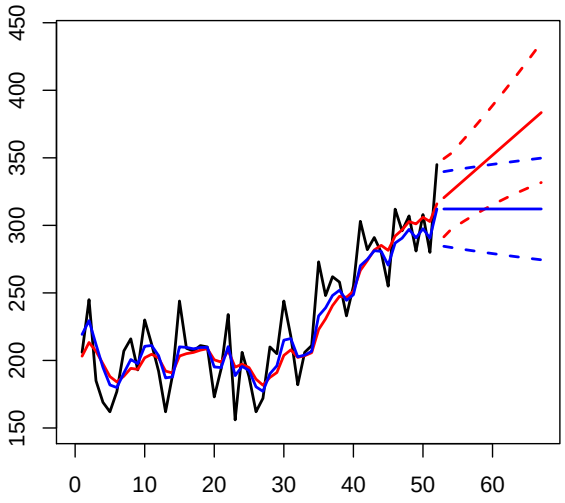
To make τ-step ahead forecast at time n, we compute

$$\hat{Y}_{n+\tau} = L_n + \tau \cdot B_n$$

Given the value of L_n, B_n, the prediction is a linear function of τ. Its prediction interval can be calculated as

$$L_n + \tau B_n \pm z_{0.025} \cdot s \cdot \sqrt{1 + \sum_{j=1}^{\tau-1} \alpha^2(1+j\gamma)^2}, \quad s = \sqrt{\frac{SSE}{n-2}}$$

based on the normality assumptions.



Similarly, for different purposes, other forms have been used for trend corrected smoothing.

- Standard form

$$L_n = \alpha Y_n + (1-\alpha)(L_{n-1} + B_{n-1}),$$

$$B_n = \gamma(L_n - L_{n-1}) + (1-\gamma)B_{n-1}$$

- Correction form

$$L_n = L_{n-1} + B_{n-1} + \alpha(Y_n - L_{n-1} - B_{n-1})$$

$$B_n = B_{n-1} + \alpha\gamma(Y_n - L_{n-1} - B_{n-1})$$

9.3 Holt-Winters Method

If both seasonal trend and linear trend are present, we also need to consider the impact from the seasonality. Consider the model with additive seasonal variation,

$$Y_t = \beta_0 + \beta_1 t + SN_t + \epsilon_t$$

All parameters β_0, β_1, SN_t need to be updated.

$$\begin{aligned} L_n =& \alpha(Y_n - SN_{n-L}) + (1-\alpha)(L_{n-1} + B_{n-1}) \\ B_n =& \gamma(L_n - L_{n-1}) + (1-\gamma)B_{n-1} \\ SN_n =& \delta(Y_n - L_n) + (1-\delta)SN_{n-L} \end{aligned}$$

It can be noted that all three smoothing are based on the same error term E_n. The following form are simpler to implement in practice.

$$\begin{aligned} E_n =& Y_n - (L_{n-1} + B_{n-1} + SN_{n-L}) \\ L_n =& L_{n-1} + B_{n-1} + \alpha E_n \\ B_n =& B_{n-1} + \alpha\gamma E_n \\ SN_n =& SN_{n-L} + (1-\alpha)\delta E_n \end{aligned}$$

A point forecast for τ step later at n is $\hat{Y}_{n+\tau} = L_n + \tau \cdot B_n + SN_{n+\tau-kL}$ The 95% prediction interval is $\hat{Y}_{n+\tau} \pm z_{0.025} s\sqrt{c_\tau}$, where

$$c_\tau = \begin{cases} 1, & \tau = 1 \\ \left[1 + \sum_{j=1}^{\tau-1} \alpha^2(1+j\gamma)^2\right], & 2 \le \tau \le L \\ 1 + \sum_{j=1}^{\tau-1}[\alpha(1+j\gamma) + d_{j,L}(1-\alpha)\delta]^2, & L \le \tau \end{cases}$$

For models with multiplicative seasonal variation,

$$Y_t = (\beta_0 + \beta_1 t) \cdot SN_t \cdot \epsilon_t$$

Changes can be made in the smoothing

$$\begin{aligned} L_n &= \alpha(Y_n/SN_{n-L}) + (1-\alpha)(L_{n-1} + B_{n-1}) \\ B_n &= \gamma(L_n - L_{n-1}) + (1-\gamma)B_{n-1} \\ SN_n &= \delta(Y_n/L_n) + (1-\delta)SN_{n-L} \end{aligned}$$

A point forecast for τ step later at n is $\hat{Y}_{n+\tau} = (L_n + \tau \cdot B_n) \cdot SN_{n+\tau-kL}$. The 95% prediction interval is $\hat{Y}_{n+\tau} \pm z_{0.025} s_r \cdot \sqrt{c_\tau} \cdot SN_{n+\tau-L}$, where

$$s_r = \sqrt{\frac{\sum_{i=1}^{n} [\frac{Y_i - \hat{Y}_i(i-1)}{\hat{Y}_i(i-1)}]^2}{n-3}}$$

10 ARMA Time Series Model

In this chapter, we are discussing a forecasting methodology for stationary time series. It finds the best fit of a time series to past values in order to make forecasts. The methodology is named after George Box and Gwilym Jenkins, called **Box-Jenkins methodology**. They include two major types of time-series models: i) Autoregressive (AR) Models; ii) Moving Average (MA) Models. Combining them together, Box-Jenkins methods are also known as Autoregressive Moving Average (ARMA) models or Autoregressive Integrated Moving Average (ARIMA) models. Details will be discussed later.

10.1 Stationary

Definition 10.1 *A time series is stationary if its statistical properties, e.g. mean and variance, are essentially constant through time.*

In particular, if a series ϵ_t has zero mean, and constant variance σ^2, and ϵ_i and ϵ_j are uncorrelated for any $i \neq j$, the sequence is called white noise sequence.

When data is not stationary, as in many examples, transformation might be needed. Typical transformation includes differencing the time series in different degree, e.g.,

- First order difference

$$z_t = y_t - y_{t-1}, \quad t = 2, 3, \cdots, n$$

- Second order difference

$$z_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}, \ t = 3, \cdots, n$$

- Seasonal adjustment

10.2 ACF and PACF

Recall that correlation between two random variable X and Y is used to measure the strength of their linear relationship:

$$r = \frac{Cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}} \approx \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sqrt{[n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2][n\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2]}}$$

when n pairs of observations of them are given.

Similarly, to measure the dependence between past variables and current variables, we can compute the correlation between Z_t and Z_{t-k}. Since the correlation is defined for the same random variable at different times, it is called autocorrelation. The lag k measures the correlation between observations apart from each other in k steps

$$\rho_k = \frac{Cov(Z_t, Z_{t+k})}{var(z_t)}.$$

The definition is only meaningful when the time series is stationary, such that Z_t and Z_{t-k} have the same mean and variance. It can be estimated from the data

$$\hat{\rho}_k = \frac{\sum_{t=b}^{n-k}(z_t - \bar{z})(z_{t+k} - \bar{z})/(n-k-b+1)}{\sum_{t=b}^{n}(z_t - \bar{z})^2/(n-b+1)},$$

where $\bar{z} = \sum_{t=b}^n z_t/(n-b+1)$ is the sample mean of the series.

Like other statistics based on the data, $\hat{\rho}_k$ is random, and has corresponding standard error. This standard error can be used to assess whether the autocorrelation is *statistically significant* from 0.

$$SE(\hat{\rho}_k) = \sqrt{\frac{1 + 2\sum_{j=1}^{k-1}\rho_j^2}{n-b+1}}$$

In particular, for $\hat{\rho}_1$, we have $SE(\hat{\rho}_1) = \sqrt{1/(n-b+1)}$. Similar to the test of regression coefficients, if $|\hat{\rho}_k/SE(\hat{\rho}_k)| > t_{n-p-1,\alpha/2}$, we can claim that the autocorrelation is significant at level α at lag k.

Similar to autocorrelation, a closed related concept is Partial Autocorrelation Function. The partial autocorrelation at lag k may be viewed as the autocorrelation of time series observations separated by a lag of k time units with the effect of the intervening observations eliminated. It can be computed based on the autocorrelation function. In particular,

$$\begin{aligned} r_{11} =& \rho_1 \\ r_{kk} =& \frac{\rho_k - \sum_{j=1}^{k-1} r_{k-1,j}\cdot\rho_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j}\cdot\rho_j}, \quad k = 2,3,\cdots \end{aligned}$$

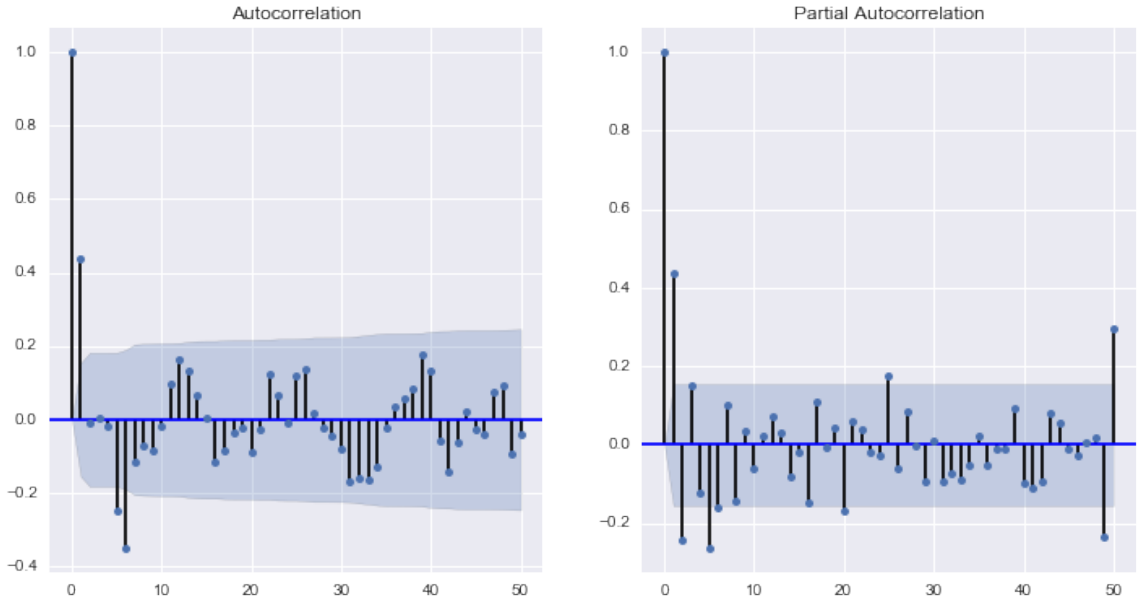
where $r_{k,j} = r_{k-1,j} - r_{kk} \cdot r_{k-1,k-j}$ for $j = 1, 2, \cdots, k-1$.

Similar to the sample autocorrelation, we can obtain the sample PAC based on the time series observations. The standard error of the SPAC can be obtained

$$SE(\hat{r}_{kk}) = \frac{1}{\sqrt{n-b+1}},$$

which is constant regardless of the choice of k.

Both ACF and PACF are characteristics of dependence among samples at different lags. They are often used jointly to determine the dependance nature of the sequence.



10.3 ARMA model

AR and MA are two fundamental building blocks of the Box-Jenkins methods.

(I) Moving Average (MA) models

The MA model assumes the time series are generated by the moving average of white noise. As a result, the autocorrelation of the data is caused by the overlapping in computing the moving average. In general, a MA model with order q is specified as

$$z_t = \delta + \epsilon_t + \theta_1 \cdot \epsilon_{t-1} + \theta_2 \cdot \epsilon_{t-2} + \cdots + \theta_q \cdot \epsilon_{t-q} \tag{10.1}$$

where ϵ_t are white noise (or iid normal), and cannot be directly observed. The process has

$$E(z_t) = \mu, \quad var(z_t) = \sigma^2(1 + \sum_{j=1}^{q} \theta_j^2)$$

By the construction, we can observe that the autocorrelation at lag $k > q$ should be zero as the moving average windows do not overlap any more. More specifically, it has the following (theoretical) autocorrelation function

$$\begin{aligned} \rho_k &= \sigma^2 \sum_{j=0}^{q-k} \theta_j \theta_{j+k} / var(z_t), \quad k \leq q, \\ &= 0, \quad k > q \end{aligned} \tag{10.2}$$

where for notation simplicity, θ_0 is defined to be 1.

Example:

(a) MA(1) model

$$z_t = \delta + \epsilon_t + \theta_1 \cdot \epsilon_{t-1},$$

- $\rho_1 = \theta_1/(1+\theta_1^2), \rho_k = 0, k \geq 2$
- AC cuts off after lag 1

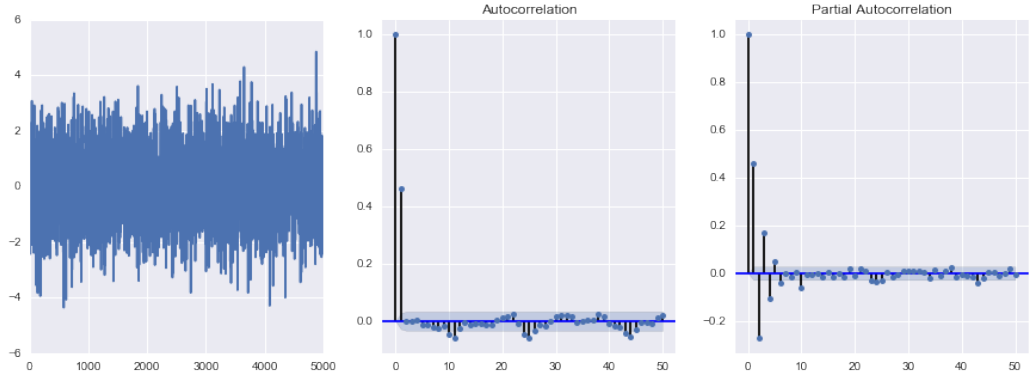
(b) MA(2) model

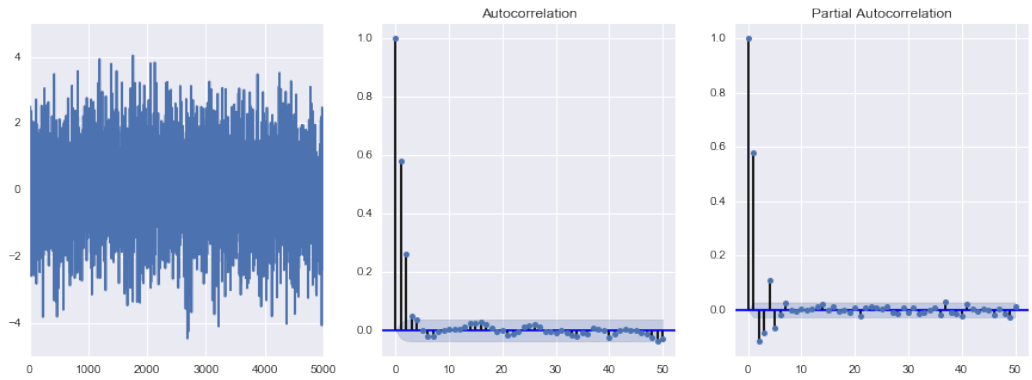
$$z_t = \delta + \epsilon_t + \theta_1 \cdot \epsilon_{t-1} + \theta_2 \cdot \epsilon_{t-2}$$

- The autocorrelation function

$$\rho_1 = \frac{\theta_1(1+\theta_2)}{1+\theta_1^2+\theta_2^2}, \quad \rho_2 = \frac{\theta_2}{1+\theta_1^2+\theta_2^2}, \quad \rho_k = 0, k \geq 3.$$

- AC cuts off after lag 2





(II) Autoregressive (AR) models

The AR model assumes the time series are generated by explicitly regress on its previous values. As a result, the autocorrelation of the data is caused by the direct dependence on previous data. In general, a AR model with order p is specified as

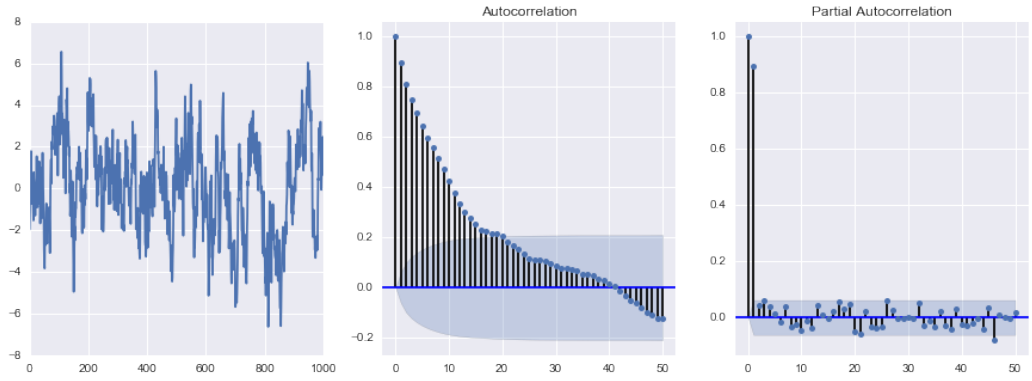
$$z_t = \delta + \phi_1 \cdot z_{t-1} + \phi_2 \cdot z_{t-2} + \cdots + \phi_p \cdot z_{t-p} + \epsilon_t \tag{10.3}$$

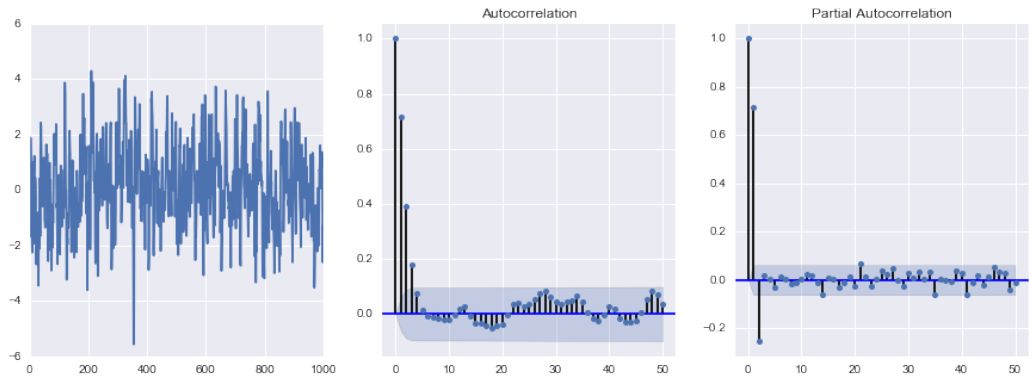
where ϵ_t are white noise (or iid normal). Since z_t can be directly observed, the AR model can be obtained by multiple linear regression, setting z_t as the response, and $z_{t-1}, z_{t-2}, \cdots, z_{t-p}$ as covariates.

Because of the explicit dependence on the past data, the autocorrelation function has a recursive formula

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \cdots + \phi_p \rho_{k-p}$$

In general, ρ_k has an exponential behavior and cyclical patterns. In contrast, the PACF of AR(p) model cuts down to zero after lag $k = p$.





Example: Consider the AR(1) model, $z_t = \delta + \phi_1 \cdot z_{t-1} + \epsilon_t$ its PACF is

$$r_{11} = \phi_1, \quad r_{kk} = 0, k \geq 2.$$

(III) ARMA(p, q) model

The ARMA(p, q) is a combination of both components. It has the general form

$$z_t = \delta + \phi_1 \cdot z_{t-1} + \phi_2 \cdot z_{t-2} + \cdots + \phi_p \cdot z_{t-p} + \epsilon_t + +\theta_1 \cdot \epsilon_{t-1} + \theta_2 \cdot \epsilon_{t-2} + \cdots + \theta_q \cdot \epsilon_{t-q} \quad (10.4)$$

Sometimes it is more organized to shift the z_t and ϵ_t to two sides of the equation

$$z_t - \phi_1 \cdot z_{t-1} - \phi_2 \cdot z_{t-2} - \cdots - \phi_p \cdot z_{t-p} = \delta + \epsilon_t + \theta_1 \cdot \epsilon_{t-1} + \theta_2 \cdot \epsilon_{t-2} + \cdots + \theta_q \cdot \epsilon_{t-q}$$

Introducing the backshift operator, $\mathcal{B}$, which has the effect $\mathcal{B}z_t = z_{t-1}$, and $\mathcal{B}^k z_t = z_{t-k}$, then we have

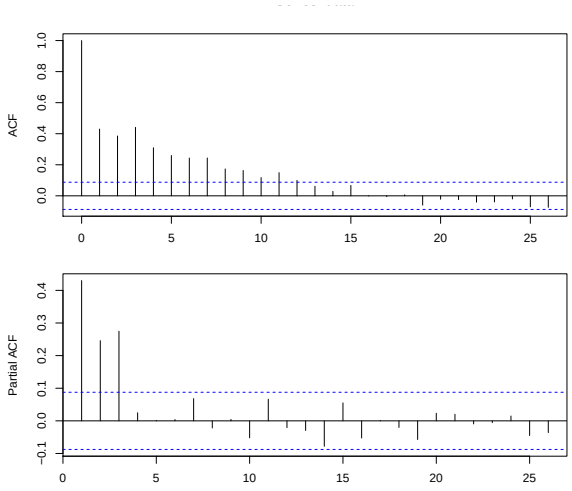
$$(1 - \phi_1 \mathcal{B} - \phi_2 \mathcal{B}^2 - \cdots - \phi_p \mathcal{B}^p) z_t = \delta + (1 + \theta_1 \mathcal{B} + \theta_2 \mathcal{B}^2 + \cdots + \theta_q \mathcal{B}^q) \epsilon_t.$$

It can be viewed as polynomial of $\mathcal{B}$ of the coefficient process. In fact, the polynomials of $\mathcal{B}$ play a central role in determining the properties of the ARMA process.

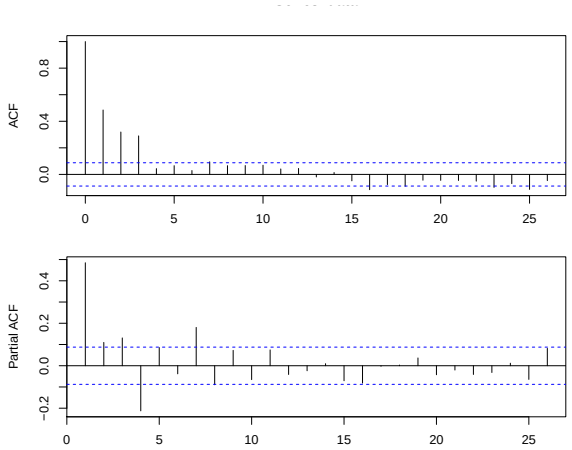
10.4 ARMA model

As a preliminary identification of the model types and orders, we can use the following observations:

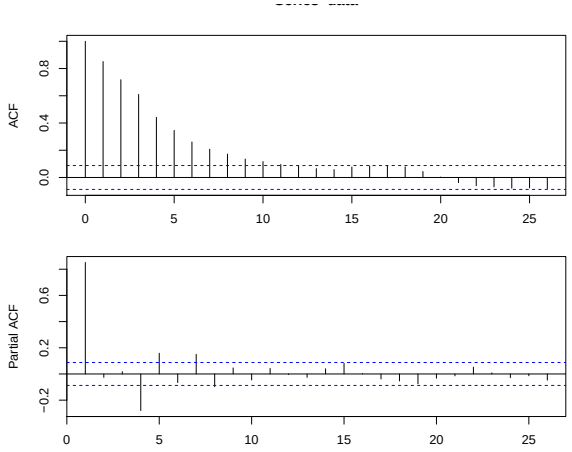
- AR(p) model typically has autocorrelation function dying down, and the partial autocorrelation function cuts off at lag p.



- MA(q) model: the autocorrelation function cuts off at lag q, and partial autocorrelation dies down.



- ARMA(p, q): both autocorrelation and partial autocorrelation dies down.



10.4.1 Link to other models

ARMA model has close link to other time series analysis techniques.

1. **Simple smoothing**: The forecasting with *simple* exponential smoothing is equivalent to forecasting with
$$z_t = \epsilon_t - \theta_1 \epsilon_{t-1}, \quad where\ z_t = y_t - y_{t-1}$$
In this case, the smoothing constant $\alpha = 1 - \theta_1$. The previous forecast errors are used to adjust current forecast.

2. **Holt's trend corrected smoothing**: The forecasting with *trend corrected* exponential smoothing is equivalent to forecasting with
$$z_t = \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2}, \quad where\ z_t = y_t - 2y_{t-1} + y_{t-2}$$
In this case, the smoothing constant
$$\theta_1 = 2 - \alpha - \gamma, \quad \theta_2 = \alpha - 1.$$

3. **Regression on time**: In time series regression, we often assume $y_t = TR_t + SN_t + \xi_t$ and ξ_t are *independent* and normally distributed. In many applications, we may find ξ_t are correlated. In such cases, we can use Box-Jenkins model to model correlated ξ_t, and combine them together
$$y_t = TR_t + SN_t + \xi_t, \quad \xi_i \sim \text{ARMA}(p, q)$$

For notation simplicity, the Box-Jenkins methods often use the following notation

$$y_t \sim \text{ARIMA}(p, d, q)$$

where

- p: is the order of the AR terms
- q: is the order of the MA terms
- d: is the order of differencing

The notation implies

$$z_t = \delta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_p z_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_p \epsilon_{t-q}$$

where $z_t = (1 - \mathcal{B})^d y_t$ is dth order differencing.

Examples:

- ARIMA(1,0,0) becomes AR(1)

$$y_t = \mu + \phi_1 y_{t-1} + \epsilon_t$$

- ARIMA(0,0,2) becomes MA(2)

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

- ARIMA(0,1,1) becomes IMA(1,1)

$$y_t - y_{t-1} = \mu + \epsilon_t + \theta_1 \epsilon_{t-1}$$

10.4.2 Model Constraints

The parameters of the ARIMA models need to satisfy a few constraints to make the model meaningful and easy to interpret. Among them, the following two are most crucial.

- **Stationary (causal) condition**: the roots of the following equation must satisfy $|z| > 1$

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

- **Invertible condition**: the roots of the following equation must satisfy $|z| > 1$

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

These constraints are in place for both true model parameters and estimated model parameters.

10.4.3 Model Prediction

In general ARMA model, τ-step forecasting can be made at time t

$$\hat{y}_{t+\tau} = \delta + \phi_1 \hat{y}_{t+\tau-1} + \cdots + \phi_p \hat{y}_{t+\tau-p} + \hat{\epsilon}_{t+\tau} + \theta_1 \hat{\epsilon}_{t+\tau-1} + \cdots + \theta_q \hat{\epsilon}_{t+\tau-q}$$

If $y_{t+\tau-i}$ is observed, we use the observed values ($\hat{y}_{t+\tau-i} = y_{t+\tau-i}, \quad \tau \leq i$), otherwise, the forecasted values at previous steps are used. In contrast, if $\epsilon_{t+\tau-i}$ is beyond current time step t, it is set to 0 ($\hat{\epsilon}_{t+\tau-i} = 0, \quad \tau \geq i$). Otherwise, $\hat{\epsilon}_i$ is estimated by the ith step prediction error.

1. AR(p) model: Given the data up to time $t-1$, the forecast at t is naturally

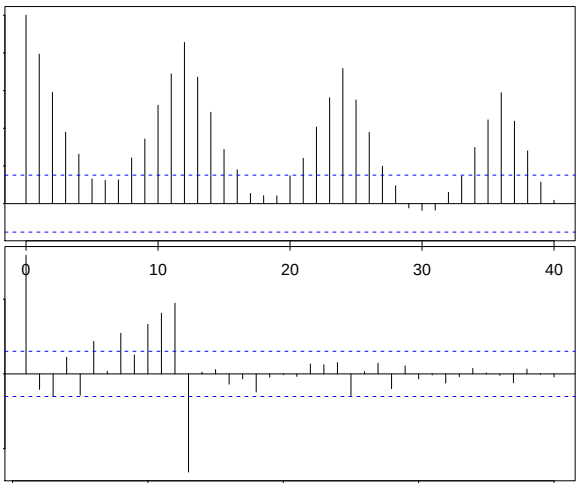
$$\hat{y}_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$$

with forecast error $e_t = y_t - \hat{y}_t$, and the sum of squared error $\sum_{i=p+1}^{n} (y_t - \hat{y}_t)^2$.

2. MA(1) model: $y_t = \delta + \epsilon_t - \theta_1 \epsilon_{t-1}$. The values of ϵ_t are observable. Using previous forecast error to estimate ϵ_{t-1}, , $\hat{\epsilon}_{t-1} = y_{t-1} - \hat{y}_{t-1}$. Then the forecast values can be computed by $\hat{y}_t = \delta - \theta_1 \hat{\epsilon}_{t-1}$, with the sum of squared forecast error as $SSE = \sum_{i=2}^{n} (y_t - \hat{y}_t)^2 = \sum_{i=2}^{n} \hat{\epsilon}_t^2$

10.5 Seasonal ARMA model

When the time series has seasonal effects, the autocorrelation often shows the seasonality, as demonstrated below

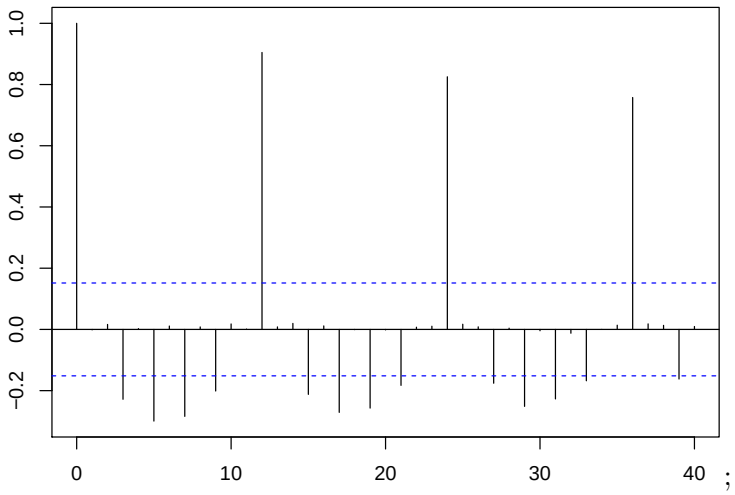


This is also a sign of non-stationarity. In general, we need to check the SAC at two different levels. At **nonseasonal level**, the SAC at lags ranging from 1 to $L-3$, is used to indicate the stationarity (whether trend exists), similar to non-seasonal ARMA models. At **seasonal level**, the SAC at lags around $L, 2L, 3L, \cdots$ indicate the correlation between the same season in different periods. Both levels should die down or cut off quickly to indicate stationarity.

If the time series is nonstationary, we can use the first order or higher order differencing to make them stationary.

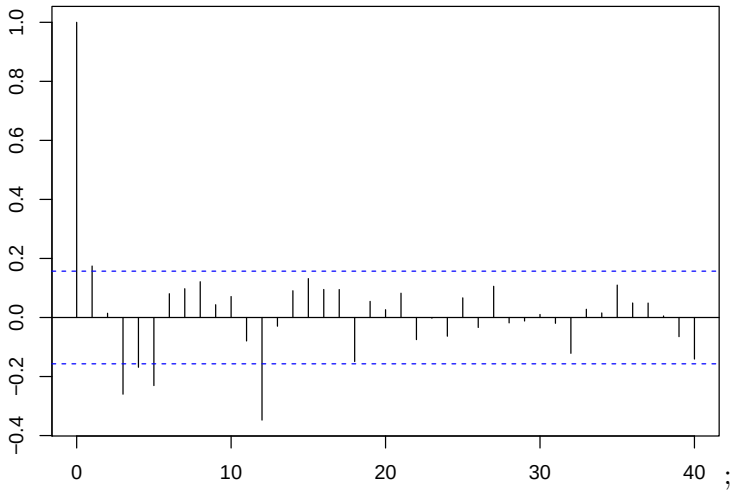
1. First *regular* differencing: $z_t = y_t - y_{t-1}$
2. First *seasonal* differencing: $z_t = y_t - y_{t-L}$
3. First regular and seasonal differencing: $z_t = y_t - y_{t-1} - (y_{t-L} - y_{t-L-1})$

Back to the picture above, we can try these three differencing ways to make the series stationary. After regular differencing, the SAC is



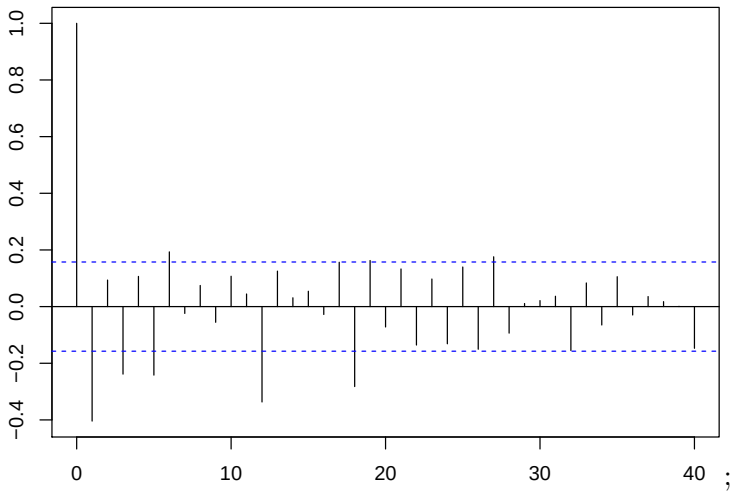
;

After seasonal differencing, the SAC is



;

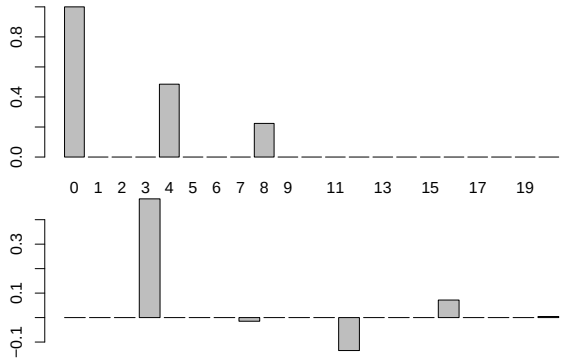
After regular and seasonal differencing, the SAC is



;

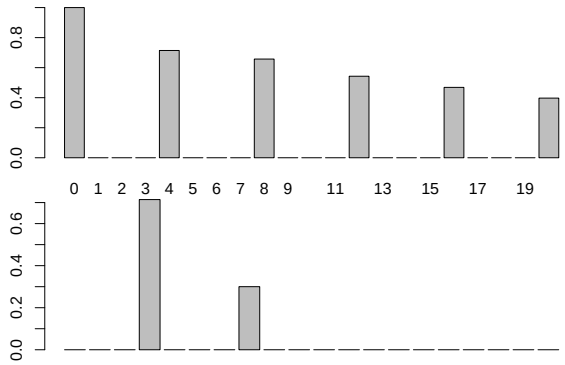
Similar to ARMA model, we can define the model at seasonal level. For seasonal models with period L, their counter part can be defined as

1. Seasonal moving average model of order Q: $z_t = \delta + \epsilon_t + \theta_1\epsilon_{t-L} + \theta_2\epsilon_{t-2L} + \cdots + \theta_Q\epsilon_{t-QL}$



$$z_t = \epsilon_t - 0.5\epsilon_{t-4} - 0.3\epsilon_{t-8}$$

2. Seasonal moving average model of order P: $z_t = \delta + \phi_1 z_{t-L} + \phi_2 z_{t-2L} + \cdots + \phi_P z_{t-PL} + \epsilon_t$

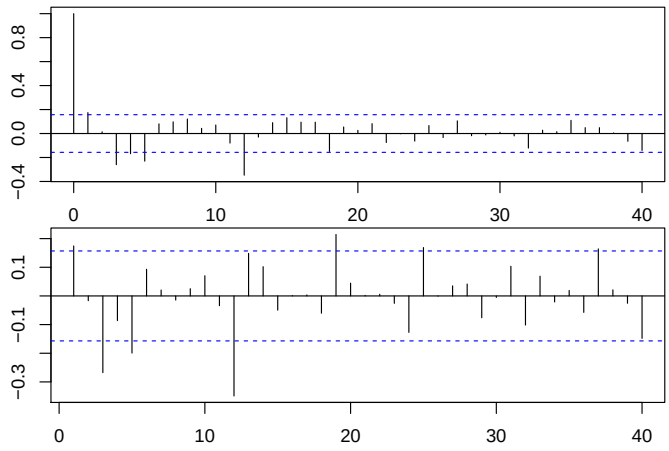


$$z_t = 0.5z_{t-4} + 0.3z_{t-8} + \epsilon_t$$

To tentatively specify the seasonal Box-Jenkins model

- Use the behavior of SAC and SPAC at **nonseasonal** level to identify a **nonseasonal** model
- Use the behavior of SAC and SPAC at **seasonal** level to identify a **seasonal** model
- Combine the model identified in the previous two steps

Example:
After first seasonal differencing, the data is stationary



The SPAC cuts off after lag 5 with spikes at 1,3,5,which indicates$z_t = \delta + \phi_1 z_{t-1} + \phi_3 z_{t-3} + \phi_5 z_{t-5} + \epsilon_t$. The SAC cuts off after lag 1 at **seasonal** level,which indicates $z_t = \delta + \epsilon_t - \theta_1 \epsilon_{t-12}$.Combine the model, we can have the final model:$z_t = \delta + \phi_1 z_{t-1} + \phi_3 z_{t-3} + \phi_5 z_{t-5} + \epsilon_t - \theta_1 \epsilon_{t-12}$

To make point forecasting, we first forecast on the z_t, and then y_t. Firstly, use Box-Jenkins model to obtain the point and interval forecast for $z_{t+\tau}$: $\hat{z}_{t+\tau} = \delta + \phi_1 z_{t+\tau-1} + \phi_3 z_{t+\tau-3} + \phi_5 z_{t+\tau-5} - \theta_1 \hat{\epsilon}_{t+\tau-12}$. Then,since first order seasonal differencing is used, $y_{t+\tau} = y_{t+\tau-12} + \hat{z}_{t+\tau}$.

When both seasonal and non-seasonal model have the same type of model (both are AR or both are MA), some new multiplicative terms are needed

$$(1 - a_1\mathcal{B} - a_2\mathcal{B}^2 - \cdots - a_p\mathcal{B}^p)(1 - \phi_1\mathcal{B}^L - \cdots - \phi_P\mathcal{B}^{PL})y_t = (1 - b_1\mathcal{B} - b_2\mathcal{B}^2 - \cdots - b_q\mathcal{B}^q)(1 - \theta_1\mathcal{B}^L - \cdots - \theta_Q\mathcal{B}^{QL})\epsilon_t \tag{10.5}$$

There is no problem in specifying the model, but we need to be careful in parameter estimation and forecasting.

ARIMA model with seasonal effects is also short for SARIMA

$$y_t \sim \text{SARIMA}(p, d, q) \times (L, P, D, Q)$$

- L: the period of the seasonal effect
- P: the order of AR part at seasonal level
- Q: the order of MA part at seasonal level
- D: the order of differencing at seasonal level

Example: Consider $SARIMA(0, 1, 1) \times (12, 0, 1, 1)$ model. It has period $L = 12$. Using Lag operator, we have

$$(1 - \mathcal{B})(1 - \mathcal{B}^{12})y_t = (1 + \theta_1\mathcal{B})(1 + a_1\mathcal{B}^{12})\epsilon_t$$

It is equivalent to

$$(1-\mathcal{B}-\mathcal{B}^{12}+\mathcal{B}^{13})y_t =(1+\theta_1\mathcal{B}+a_1\mathcal{B}^{12}+\theta_1 a_1\mathcal{B}^{13})\epsilon_t$$

When multiplicative seasonal effects are present, it can be first transformed to additive effects, and then use the SARIMA model to model the variability over time.

10.6 Model Estimation

Given the model structure and order, there are generally three types of estimation methods. The first one matches the sample ACF and sample PACF with the theoretical ACF and PACF (in the function of parameters), and get their estimated values. The second category is the least square method, and the third category is maximum likelihood estimation.

10.6.1 Matching TAC with SAC(Moment Method)

Given the tentatively determined model, its TAC is often available. For example, the MA(1) model $y_t = \delta + \epsilon_t + \theta_1\epsilon_{t-1}$ has TAC

$$\rho_1 = \frac{\theta_1}{1+\theta_1^2}, \quad \rho_k = 0, \ \ \forall k > 1$$

By equating TAC with SAC ($\rho_1 = r_1$), we can get

$$\hat{\theta}_1 = \frac{1 \pm \sqrt{1-4r_1^2}}{2r_1}.$$

Similar idea can be applied to other models with relatively low orders, as long as the TAC is analytically available.

(a) MA(2) model

$$\rho_1 = \frac{\theta_1(1+\theta_2)}{1+\theta_1^2+\theta_2^2}, \ \rho_2 = \frac{\theta_2}{1+\theta_1^2+\theta_2^2}, \ \rho_k = 0, k > 2$$

(b) AR(1) model

$$\rho_k = \phi_1^k, \quad k \geq 1$$

(c) AR(2) model

$$\rho_1 = \frac{\phi_1}{1-\phi_2}, \ \rho_2 = \frac{\phi_1^2}{1-\phi_2} + \phi_2, \ \rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2}, k \geq 3$$

(d) ARMA(1,1) model

$$\rho_1 = \frac{(1+\phi_1\theta_1)(\phi_1+\theta_1)}{1+\theta_1^2+2\phi_1\theta_1}, \quad \rho_k = \phi_1\rho_{k-1}, \ k \geq 2$$

By matching the theoretical AC with sample AC, we can calculate the parameters of interest for corresponding model. This method can be applied to more complex model as well, and is known as Yule-Walker estimation in general. It is simple, and does not require the original raw data. However, the estimation accuracy is not the best. When multiple solutions exists, it is important to check whether these solutions can satisfy the stationary and invertible conditions of the model.

10.6.2 Least square and MLE

These two methods are not elaborated, and can take full advantage of the entire dataset. The least square method tries to find out the parameters such that the sum of squared forecast error is minimized. It is similar to the estimation in conventional regression analysis. In the context of ARMA models, calculating the forecast errors is straightforward for AR(p) model. However, it needs more care when MA component is involved, e.g., MA(q) and ARMA(p, q) model.

The maximum likelihood estimation requires the joint distribution of all observations (typically multivariate normal). It generally provides more accurate estimation, but also is more complex. Software can be used to obtain such estimates.

10.6.3 Model Diagnostics

Similar to regression models, a good way to check the adequacy of an Box-Jenkins model is to analyze the residuals

$$e_t = y_t - \hat{y}_t.$$

In particular, we can plot the SAC and SPAC for the residuals to check whether the model is adequate. These plots are often named as RSAC and RSPAC, respectively, for short. If the model is adequate, the error should be uncorrelated, and the RSAC be small. Detailed plot of RSAC or RSPAC can be used to improve the model as well. In addition, we can also use some statistic to quantify the dependence of the residuals.

One of such statistic is the Ljung-Box Statistic. The statistic is computed as

$$Q^* = (n-d)(n-d+2)\sum_{l=1}^{K}(n-d-l)^{-1}r_l^2(\epsilon),$$

where n is the sample size, d is the number of differencing, $r_l(\epsilon)$ is the SAC of the residual at lag l, K is some number indicating the range of interests. If $Q^* > \chi^2_{\alpha,K-n_C}$, the residuals are correlated, i.e., the model is inadequate. In practice, multiple K *(e.g.=6,12,18,24) can be used to check the correlation of the residuals.

11 Spatial Data Forecasting

11.1 Spatial Data

Spatial data comes from a myriad of fields, which lead to various spatial data types. A general and useful classification of spatial data is provided by Cressie (1993, pp. 8-13) based on the nature of the spatial domain under study.

Following Cressie (1993), let $\boldsymbol{s} \in \mathcal{R}^d$ be a generic location in a d-dimensional Euclidean space and $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}^d\}$ be a spatial random function, Z denote the attribute we are interested in. The three spatial data types are: **lattice data**, **geostatistical data**, and **point processes**.

- Lattice (Areal/Regional) data: the domain D under study is discrete. Data can be exhaustively observed at fixed locations that can be enumerated. Locations can be ZIP codes, neighborhoods, provinces etc. Data in most of cases are spatially aggregated. Eg: the unemployment rate by states, crime data by counties, average housing prices by provinces.
- Geostatistical data: the domain under study is a fixed continuous set D. Eg: the level of a pollutant in a city, the precipitation or air temperature values in a country. They can have value at any point in D.
- Point processes (Point patterns): the attribute under study is the location of events (observations). Therefore, the domain D is random. The observation is not necessarily labeled and the interest lies mainly in where the events occur. Eg: the location of trees in a forest, the location of nests in a breeding colony of birds.

Figure 1 shows some examples of the spatial data in each category. The main goal of different spatial data types can be different:

- Lattice data analysis: smoothing and clustering acquire special importance. It is of interest to describe how the value of interest at one location depends on nearby values, and whether this dependence is direction dependent.
- Geostatistical data: to predict value of interest at unobserved locations across the entire domain of interest. An exhaustive observation of the spatial process is not possible. Observations are only made at a small subset of locations.
- Point processes analysis: to determine if the location of events tends to exhibit a systematic pattern over the area under study or, on the contrary, they are randomly distributed.

11.2 Lattice Data Analysis

As indicated earlier, for lattice data, the key interest is to identify their spatial dependency pattern. Before analyzing spatial dependency, it is essential to test the existeance of spatial dependency.

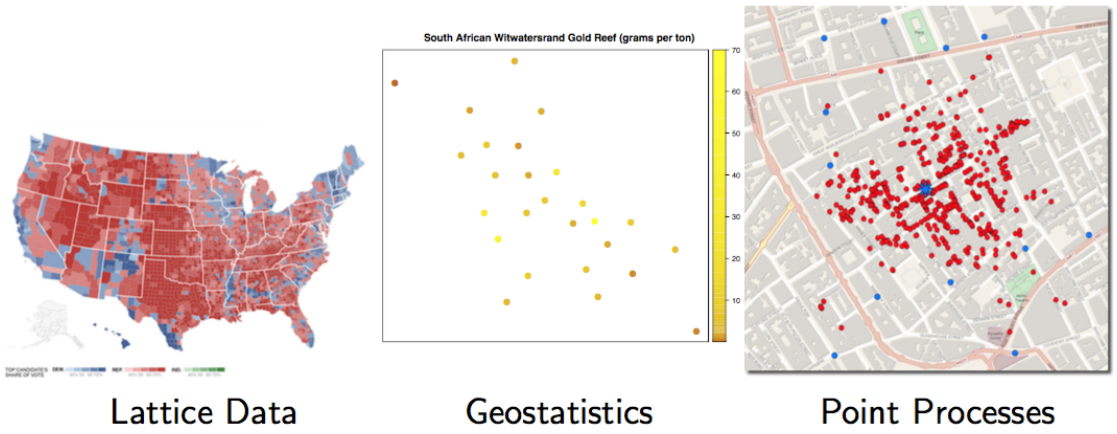


Figure 1: Three spatial data types

11.2.1 Moran's I to Test Dependency

Moran's I statistic is a widely used measure for spatial correlation (dependencies):

$$I = \frac{N}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}.$$

Here N is the number of spatial units indexed by i or j. x is the variable of interest. $\bar{x}$ is the mean of $x_i, i = 1, \cdots, N$. w_{ij} is the (i, j)th element of a spatial weight matrix W with $w_{ii} = 0$ and $S_0 = \sum_{i,j} w_{ij}$. The design of spatial weight matrix W can be:

- $w_{ij} = 1$ if zone i and zone j are neighbors. $w_{ij} = 0$ otherwise.
- $w_{ij} = 1$ if zone j is one of the k nearest neighbors of zone i. $w_{ij} = 0$ otherwise.
- w_{ij} is set based on a decay function of distance between zone i and zone j. An example can be Gaussian kernel:

  $$w_{ij} = \frac{1}{\sqrt{2\pi}h} \exp[-d_{ij}^2/(2h^2)]$$

  with d_{ij} being the distance between zone j and j, h being the bandwidth parameter to be tuned.

To test the spatial correlation, we can formualte the following hypothesis testing.

H_0: no spatial correlation; H_1: spatial correlation exists.

The H_0 indicates the spatial randomness. The expected value of Moran's I under the null hypothesis that there is no spatial correlation is $\mathbb{E}(I) = -1/(N-1)$. With large sample sizes, the expected value approaches zero. I usually range from 1 to +1. Positive I indicates positive spatial correlation, while negative I indicates negative spatial correlation. Values significantly deviate from $-1/(N-1)$

indicate spatial correlation. The variance of the statistic under the null (assuming each value is equally likely to occur at any location) is:

$$\text{Var}(I) = \frac{NS_4 - S_3S_5}{(N-1)(N-2)(N-3)S_0^2} - (\mathbb{E}(I))^2$$

where $S_1 = \sum_i \sum_j (w_{ij} + w_{ji})^2/2$; $S_2 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$;

$$S_3 = \frac{N^{-1}\sum_i (x_i - \bar{x})^4}{(N^{-1}\sum_i (x_i - \bar{x})^2)^2},$$

$S_4 = (N^2 - 3N + 3)S_1 - NS_2 + 3S_0^2$; $S_5 = (N^2 - N)S_1 - 2NS_2 + 6S_0^2$.

Alternatively, using permutation, we can obtain a reference distribution for the statistic under the null hypothesis. In each random permutation of the data across locations, we compute the statistic I. Denote M as the total number of permulations, R the number of permutation from which the computed Morans I is equal to or more extreme than the original statistic I_0. From the reference distributions from M permutation, we can calculate the p-value

$$p = \frac{R+1}{M+1}.$$

Small p value indicates existence of significant spatial correlation.

If significant spatial dependence exists, it is beneficial to characterize the dependence structure, so that this information can be used for forecasting or anomaly detection. The following few models are different approaches to characterize the spatial dependence.

11.2.2 Spatial Autoregression

One common tool to account for spatial dependency is linear regression. The idea is to have models analogous to time series models but with spatial lags. As the simplest case, when only the spatially lagged variable is considered,

$$\boldsymbol{y} = \lambda W \boldsymbol{y} + \epsilon, \quad |\lambda| < 1 \tag{11.1}$$

where $\epsilon \sim i.i.d.N(0, \sigma_\epsilon^2 I_n)$, W is a non-stochastic standardized spatial weight matrix. Compared to the spatial weight matrix explained above, it is standardized in the sense that the elements of any row sum to one, i.e., $w_{ij}^s = w_{ij}/\sum_j w_{ij}$. When the W matrix is row-standardized and $|\lambda| < 1$, the matrix $(I - \lambda W)$ is invertible. From equation 11.1 we can have $\boldsymbol{y} = (I - \lambda W)^{-1}\boldsymbol{\epsilon}$, and:

$$\mathbb{E}(\boldsymbol{y}) = 0 \tag{11.2}$$

$$\mathbb{E}(\boldsymbol{y}\boldsymbol{y}^T) = \sigma_\epsilon^2 (I - \lambda W)^{-1}(I - \lambda W^T)^{-1} = \sigma_\epsilon^2 \Omega. \tag{11.3}$$

With normality assumption of ϵ_i, this model can also be estimated via maximum likelihood procedure. The log-likelihood can be expressed as:

$$l(\lambda, \sigma_\epsilon^2) = \text{const} - \frac{n}{2}\ln(\sigma_\epsilon^2) - \frac{1}{2}\ln|(I-\lambda W)^{-1}(I-\lambda W)^{-T}| - \frac{1}{2\sigma_\epsilon^2}\boldsymbol{y}^T[(I-\lambda W)^{-1}(I-\lambda W)^{-T}]^{-1}\boldsymbol{y}.$$

The $\hat{\lambda}, \hat{\sigma}_\epsilon^2$ that maximize the likelihood function becomes the parameter estimates.

11.2.3 Spatial Linear Regression with Exogenous Variables

In addition to the spatially lagged variable as regressors, there are also some exogenous variables that can influence the response. Define the matrix of all exogenous regressors, current and spatially lagged, as $Z = [X, WX]$ and the vector of regression parameters as $\boldsymbol{\beta} = [\boldsymbol{\beta}_{(1)}, \boldsymbol{\beta}_{(2)}]$. In presence of explanatory variables, it is also possible to test the spatial dependencies, following the procedure below:

1. Run the non-spatial regression $y_i = X_i\boldsymbol{\beta}_{(1)} + e, \quad e \sim N(0, \sigma^2)$ for every y_i.
2. Test the regression residuals for spatial correlation, using Moran's I.
3. If no significant spatial correlation exists, STOP.
4. Otherwise, use a special model which takes spatial dependencies into account.

Two commonly used models considering spatial dependecies are Spatial Lag Model (SLM) and Spatial Error Model (SEM).

Spatial Lag Model (SLM):

$$\boldsymbol{y} = \lambda W\boldsymbol{y} + Z\boldsymbol{\beta} + \boldsymbol{u}, \quad |\lambda| < 1$$

$$\boldsymbol{u}|X \approx i.i.d.N(0, \sigma_u^2 I_n)$$

In this case, a problem of endogeneity emerges in that the spatially lagged value of y is correlated with the stochastic disturbance, i.e, $\mathbb{E}[(W\boldsymbol{y})\boldsymbol{u}^T] \neq 0$. Therefore, least square can not be employed. The parameters can be estimated by maximum likelihood. Sometimes this model is also called **spatial autoregressive model (SAR)**.

Spatial Error Model (SEM) :

$$\boldsymbol{y} = Z\beta + \boldsymbol{u}$$

$$\boldsymbol{u} = \rho W\boldsymbol{u} + \boldsymbol{\epsilon}, \quad |\rho| < 1$$

Compared to SLM, SEM contains spatial dependence in the noises. Similar as before, the constraints on ρ hold for row-standardized W to make $I - \rho W$ invertible. Due to the endogeneity of the errors, i.e., $\mathbb{E}[(W\boldsymbol{y})\boldsymbol{\epsilon}+^T] \neq 0$, the least square procedure loses its optimal properties.The parameters can be estimated by maximum likelihood.

11.2.4 Generalizations

Last but not least, the models discussed above are all special cases of a general form:

$$\boldsymbol{y} = \lambda W_1 \boldsymbol{y} + Z\boldsymbol{\beta} + \boldsymbol{u}, \quad |\lambda| < 1$$

$$\boldsymbol{u} = \rho W_2 \boldsymbol{u} + \boldsymbol{\epsilon}, \quad |\rho| < 1$$

Where W_1 and W_2 are not necessarily the same. This generalized model comes with several names, e.g., spatial autocorrelation model (SAC), extended spatial Durbin model (SDM), or SARAR(1,1) (acronym for spatial autoregressive with additional autoregressive error structure).

11.3 Geostatistical Interpolation

Geostatistics deals with spatially autocorrelated data [1]. The first law of geography: "everything is related to everything else, but near things are more related than distant things."(Tobler 1970)

Figure 2 shows the precipitation surface of Switzerland as an example. Blue dots are monitoring stations with size corresponding to the amount of rainfall. The different heights of the surface and their color are associated with amount of rainfalls at each location. An essential problem is to construct a continuous surface from observations at these stations.

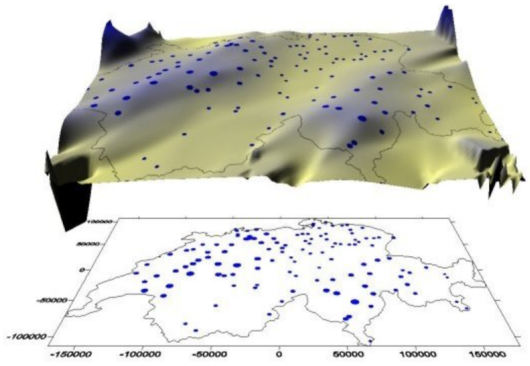


Figure 2: A precipitation surface of Switzerland.

11.3.1 Spatial Dependencies: Covariance and Semivariance

Statistically, denote $X(s)$ as the response of interest at location $s \in D$, $m(s) = \mathbb{E}X(s)$ is the mean response value. The spatial dependence between response at any two location s_i, s_j can be characterized by the following two quantities.

[1] critical reference of this section: Geographic Information Technology Training Alliance (GITTA) http://www.gitta.info/website/en/html/index.html

- Covariance: $C(s_i, s_j) = \mathbb{E}\{[X(s_i) - m(s_i)] \cdot [X(s_j) - m(s_j)]\}$,
- Semivariance: $\gamma(s_i, s_j) = \text{Var}(X(s_i) - X(s_j))/2$.

Covariance is a measure of similarity, the larger the value, the more correlated of their responses. In contrast, semivariance is calculated as a measure of dissimilarity: smaller value indicates higher dependence. They plays the pivotal role in the properties of geospatial models and their prediction accuracies. To simplify the model complexity, it is common to limit our attention to a class of stationary models:

Intrinsic stationary (IS):

- $\mathbb{E}[X(\boldsymbol{s})] = \mu$ for all $\boldsymbol{s} \in D$
- $2\gamma(\boldsymbol{h}) = \text{Var}(X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j))$, for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$, $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$ is the difference between $\boldsymbol{s}_i, \boldsymbol{s}_j$.

Second-order stationary (SOS):

- $\mathbb{E}[X(\boldsymbol{s})] = \mu$ for all $\boldsymbol{s} \in D$
- $\text{Cov}(X(\boldsymbol{s}_i), X(\boldsymbol{s}_j)) = \mathbb{E}(X(\boldsymbol{s}_i) \cdot X(\boldsymbol{s}_j)) = C(\boldsymbol{h})$, for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$, $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$.

An SOS process implies IS, which means IS is a weaker assumption. Under SOS, the relationship between semivariance and covariance is:

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}),$$

as demonstrated in Figure 3.

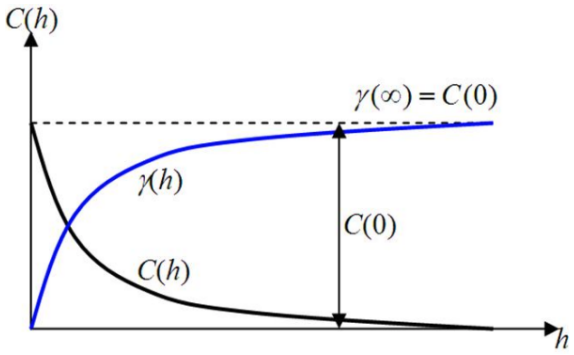


Figure 3: Relationship between covariance and semivariance under SOS.

Covariance is a more commonly seen concept in statistics. It is common to formulate geostatistic models in terms of the covariance function. Nevertheless, for estimation purposes, semivariance has more advantages:

- To estimate semivariance, no estimate of mean is required. semivariance can adapts more easily to nonstationary cases. On the contrary, covariance estimator requires the estimation of mean. When mean is unknown and needs to be estimated from sample, estimating covariance is more biased.
- Semivariance can be applied under IS, meaning that the semivariance can be defined in some cases where the covariance function cannot be defined. In particular, the semivariance may keep increasing with increasing lag, rather than leveling off, corresponding to an infinite global variance. In this case the covariance function is undefined. As a result, IS is the fundamental assumption required for Kriging instead of SOS.

Figure 4 shows an example of how semivariance works. Two datasets have similar summary statistics: 15251 points with (1) average value 100; (2) standard deviation 100; (3) median 100; (4) 10 Percentile 74; (5)90 percentile 125. However, due to different semivariance they exhibits totally different pattern (spatial structure)

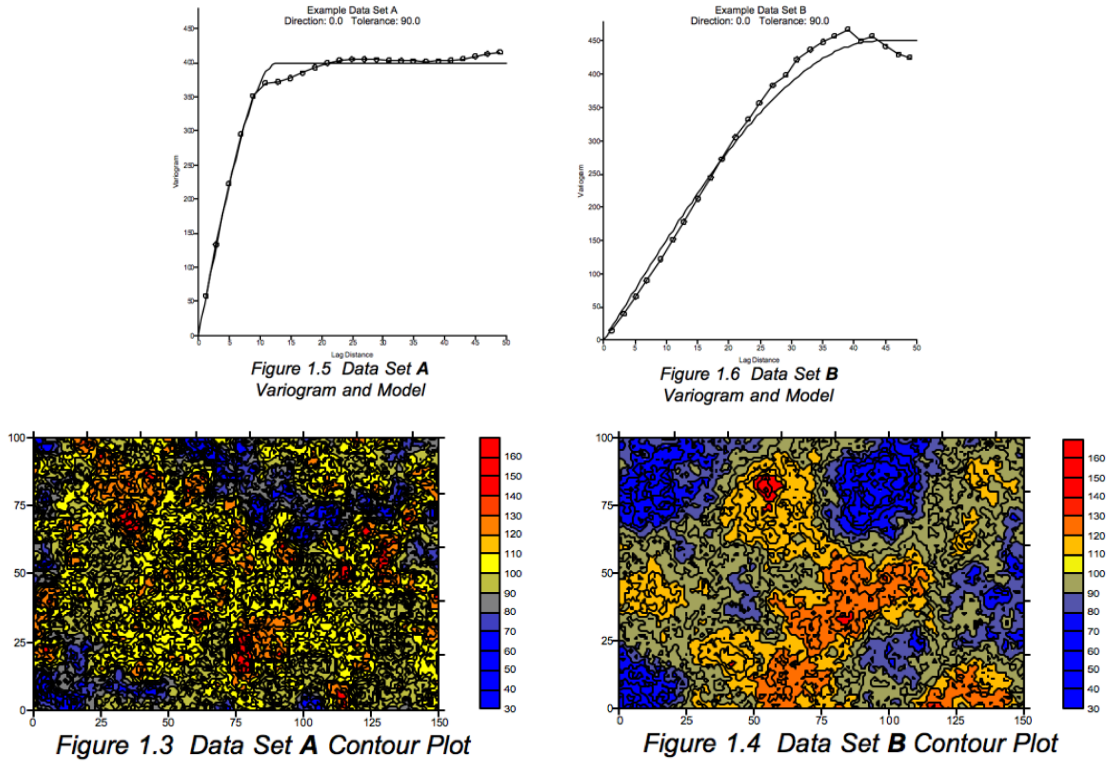


Figure 4: Different semivariance

It is note that in practice, we may further assume that the semivariance is **isotropy**, i.e., the spatial correlation is the same in all directions. There are anisotropy cases which requires further design of the semivariance formula, which we will not discuss in this note. For isotropy semivariance,

the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ completely determines their spatial correlation. As a result, we use the scalar h instead of the directional vector $\boldsymbol{h}$.

To estimate $\gamma(h)$ from the data, we can use:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i,j,\text{st} \|\boldsymbol{s}_i - \boldsymbol{s}_j\| \approx h} [X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j)]^2$$

where $N(h)$ is the number of pairs whose distances are around h. By changing the value of h, we can get the function $\hat{\gamma}(h)$, which we also refer to as **semivariogram**, as shown in Figure 5.

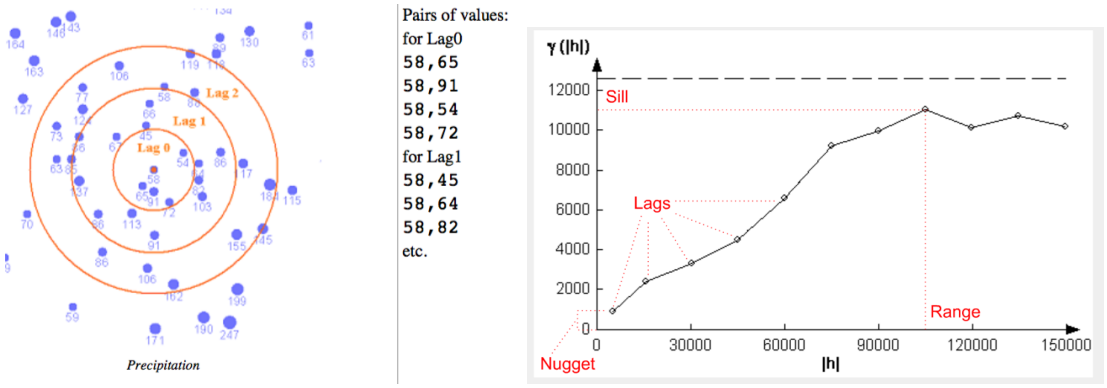


Figure 5: Left: Empirical estimation of semivariance; Right: Empirical semivariogram.

The semivariogram plot often reveals a few important points:

- Sill: The semivariance at which semivariogram levels off.
- Range: The lag distance at which the semivariogram reaches the sill value. Spatial correlation is zero beyond the range.
- Nugget: In theory the semivariogram value at the origin should be zero. If it is significantly different from zero for lags very close to zero, then this value is referred to as the nugget. (variability at distances smaller than the typical sample spacing, including measurement errors)

For modeling and prediction, we need to replace the empirical semivariogram with an acceptable parametric semivariogram model because we need to use the semivariogram values at lag distances other than empirical ones. More importantly, the semivariogram need to be non-negative definite. Let a denote the range, and c denote the sill, three most frequently used models are:

- Spherical: $\gamma(h) = c\left(1.5(h/a) - 0.5(h/a)^3\right)$ if $h \leq a$, c otherwise.
- Exponential: $\gamma(h) = c(1 - \exp(-3h/a))$
- Gaussian: $\gamma(h) = c(1 - \exp(-3h^2/a^2))$

11.3.2 Kriging as an interpolation

Given the spatial covariance (semivariance) structures of the data, we are able to predict the response at any location given the observations from a few locations. This process is also called interpolation. Interpolation algorithms predict the value at a given location as a weighted sum of data values at surrounding locations. Almost all weights are assigned according to functions that give a decreasing weight as the distance increases. Kriging is the optimal interpolation based on observed values of surrounding data points, weighted according to spatial covariance values. It also has other advantages:

- Helps to compensate for the effects of data clustering, assigning individual points within a cluster less weight than isolated data points
- Estimates standard error (kriging variance), along with estimate of the mean, which provides basis for interval forecasting .

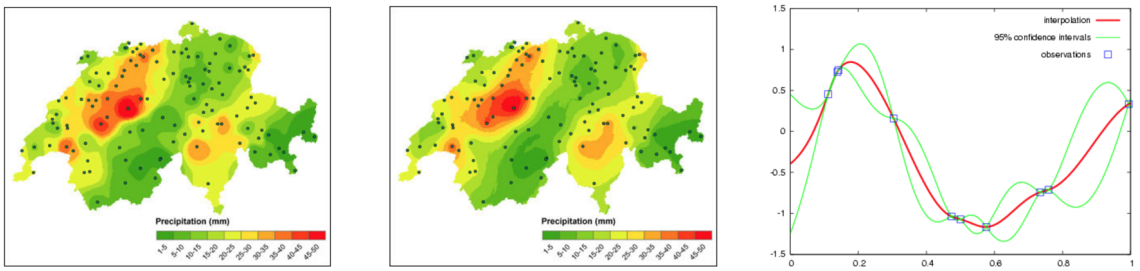


Figure 6: Left: Interpolation using inverse distance weighting; Middle: Kriging Interpolation; Right: Kriging's confidence interval.

Kriging assumes the response at location $\boldsymbol{s}$, $Z(\boldsymbol{s})$, follows $Z(\boldsymbol{s}) = m(\boldsymbol{s}) + R(\boldsymbol{s})$ in the domain D. $m(\boldsymbol{s})$ is the mean response function at location $\boldsymbol{s}$, $R(\boldsymbol{s})$ is an intrinsically stationary (IS) process. Kriging aims to predict the response value at unobserved location $\boldsymbol{s}_0$ given observations $Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2), \cdots, Z(\boldsymbol{s}_n)$. A basic form of the kriging estimator is

$$Z^*(\boldsymbol{s}_0) - m(\boldsymbol{s}_0) = \sum_{i=1}^{N(\boldsymbol{s}_0)} \lambda_i [Z(\boldsymbol{s}_i) - m(\boldsymbol{s}_i)]$$

where $N(\boldsymbol{s}_0)$ is the number of data points in the local neighborhood used for estimation of $Z^*(\boldsymbol{s}_0)$. The mathematical goal of kriging is to determine weights that minimize the variance of the estimator under the unbiasedness constraint.

$$\begin{aligned} \underset{\boldsymbol{\lambda}}{\text{minimize}} \quad & \sigma_E^2(\boldsymbol{s}) = \text{Var}\{Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\} \\ \text{subject to} \quad & \mathbb{E}\{Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\} = 0 \end{aligned}$$

Ordinary Kriging has the simplest structure for the underlying mean function, i.e., $m(\boldsymbol{s}) = \mu$. In this case, the bias is:

$$\mathbb{E}(Z^*(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)) = \left(\sum_{i=1}^{N(\boldsymbol{s}_0)} \lambda_i - 1 \right) m.$$

The unbiased estimation requires $\sum_{i=1}^{N(\boldsymbol{s}_0)} \lambda_i = 1$. The semivariance can be estimated from sample, denoted by $\gamma(h)$. The semivariance between any two observation can form a matrix Γ, where $\Gamma_{i,j} = \gamma(||\boldsymbol{s}_i - \boldsymbol{s}_j||), i, j = 1, \cdots, n$. The semivariance between $\boldsymbol{s}_0$ and existing observations can also be summarized in a vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_n]$ with $\theta_i = \gamma(||\boldsymbol{s}_0 - \boldsymbol{s}_i||)$. Minimizing the variance (uncertainty) of the prediction, we can have

$$\boldsymbol{\lambda}^* = \Gamma^{-1}(\boldsymbol{\theta} + \hat{\mu}\mathbf{1}), \quad \text{where } \hat{\mu} = \frac{1 - \mathbf{1}^T\Gamma^{-1}\boldsymbol{\theta} - 1}{\mathbf{1}^T\Gamma^{-1}\mathbf{1}},$$

and $\mathbf{1}$ is a vector of 1's.

It can be seen that the kriging weights are determined entirely by the data locations $\boldsymbol{s}$ and the covariance model, not the actual data values of $Z(\boldsymbol{s})$. Despite the simple structure and assumption of constant mean, it has been found that the ordinary Kriging is working well in many cases. It is generally unnecessary to use a complex mean function unless there are a sufficient reasons to do that.

Figure 7 is an example of the result of ordinary kriging. Compare the weight of point 5 and 6 (similar covariance and distance): point 6 is effectively screened by the nearby data point 5. Data points 5 and 6 are fairly strongly correlated with each other and 5 has a stronger correlation with the estimation point, so data point 6 is effectively ignored.

It is important to note that Kriging interpolation also comes with assumptions.

1. The underline function is from a stationary process with specified covariance function
2. If the distribution of the data is skewed, then the Kriging estimators are sensitive to a few large data values.
3. Normality of observations is not a requirement for Kriging. Nevertheless, under Gaussian assumption, Kriging is BLUE ("best linear unbiased estimator"). Kriging under Gaussian assumption is also equivalent to the the famous "Gaussian process".

Aside from the orindary Kriging, there are other variants as well.

- Simple kriging: assumes the mean response over the entire domain is a known constant: $\mathbb{E}\{Z(\boldsymbol{s})\} = m$. In this case, the constrain $\sum_{i=1}^{N(\boldsymbol{s}_0)} \lambda_i = 1$ is no longer needed.
- Universal kriging: assumes the mean response is not a constant but a linear combination of known functions: $\mathbb{E}\{Z(\boldsymbol{s})\} = \sum_{k=0}^{p} \beta_k f_k(\boldsymbol{s})$.

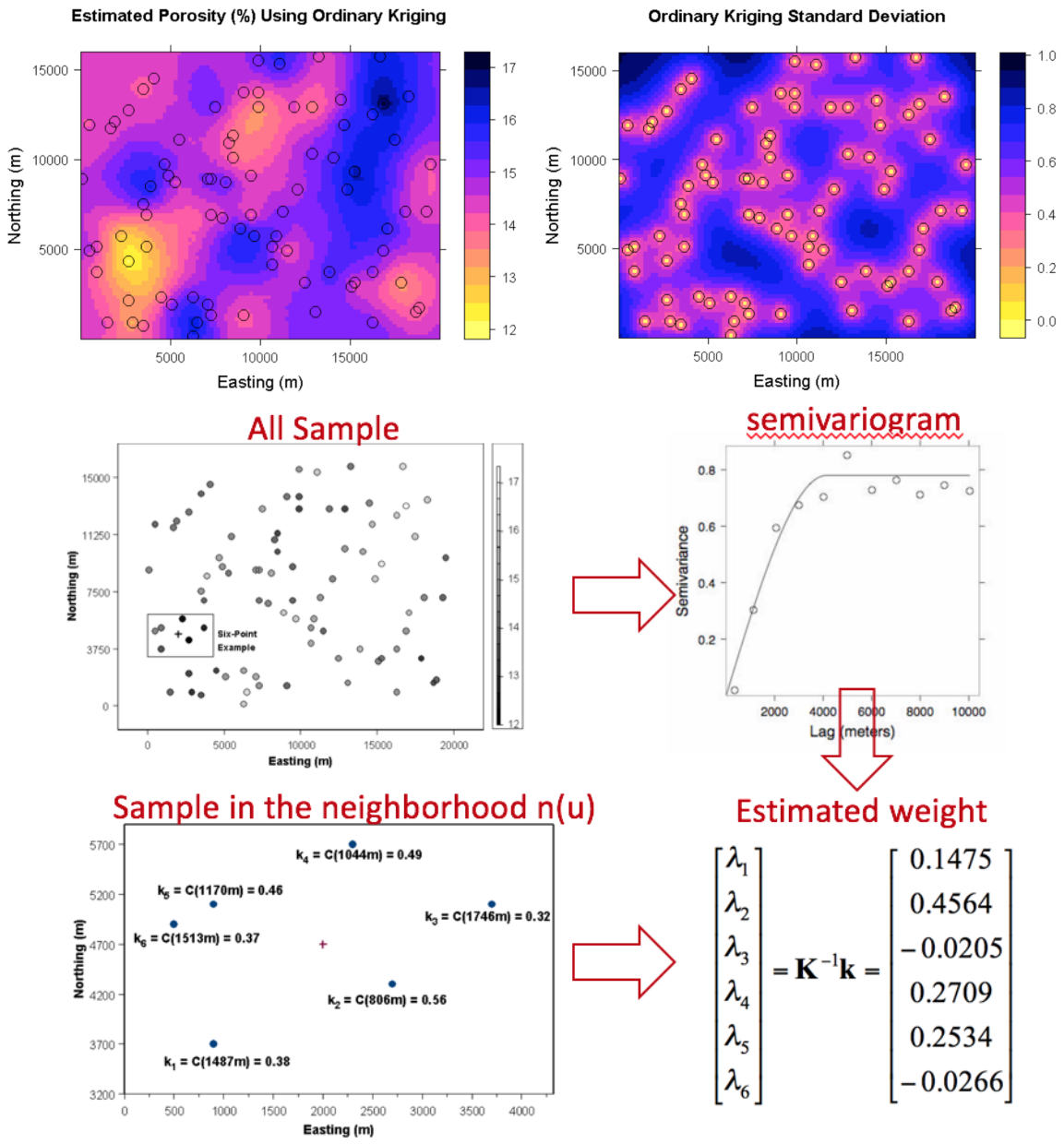


Figure 7: Example for Ordinary Kriging when $N(\boldsymbol{s}) = 6$.

- Cokriging: Kriging using information from one or more correlated variables, or multivariate kriging in general.

12 Spatial Temporal Data and Models

Spatio-temporal data refers to the data that are indexed in both space and time. Generalizing from the purely spatial setting, we use $\{Z(\boldsymbol{s}, t) : \boldsymbol{s} \in \mathcal{R}^d, t \in \mathcal{R}\}$ to denote the response as a function of both space (locations) and time. As a result, we expect that the responses are both spatially correlated and temporally correlated. Figure 8[2] provides an illustration for such data. The following list are some common examples in practice.

- If we record the Per Capita Income of every state every year, we have a collection of **spatio-temporal lattice data**. We can study how incomes of every state evolve over time.
- If we observe every hour the level of pollutant in a city at the points where the monitoring stations are located, we have a **spatio-temporal geostatistical dataset.**
- If we observe the location of bird nests every year, we have a **spatio-temporal point pattern dataset**. Now we can study whether there is complete spatio-temporal randomness or they exhibit clustering/inhibition.

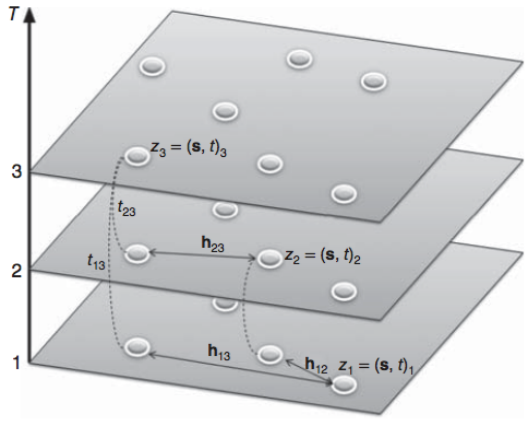


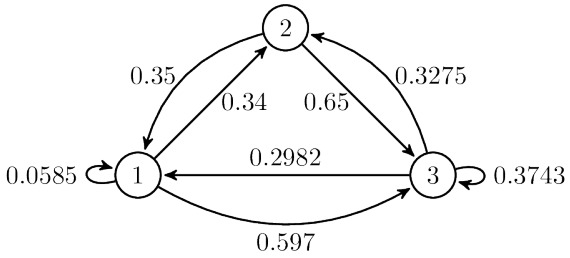
Figure 8: Spatio-Temporal Data Illustration

12.1 Spatial-temporal lattice data analysis

Spatial Markov Chain

[2]Figure cited from Fernndez-Avils, G., & Mateu, J. (2015). Spatial and spatio-temporal geostatistical modeling and kriging (Vol. 998). John Wiley & Sons.

Since locations in lattice data are discrete and finite, Markov chain can be adopted to study regional dynamics. If we divide the data into k classes and T periods, we can denote a vector $P_t = [P_{1,t}, P_{2,t}, \cdots, P_{k,t}]$ to represent the probability that the response in a region be a member of a particular class at period t, i.e., $P(Z(\boldsymbol{s}, t) \in C_k)$. To model the dynamics over time, we use the transition probability matrix M_t, whose element $m_{t,i,j}$ denotes the probability that the response currently in state i at time t ends up in state j in the next period $P(Z(\boldsymbol{s}, t+1) \in C_j | Z(\boldsymbol{s}, t) \in C_i)$. An example of the transition matrix is shown in Figure 9. If the transition probabilities do not change over time, we can drop the index t in the notations above, and we can easily get $P_{t+b} = P_t M^b$.



$$M = \begin{bmatrix} 0.0585 & 0.34 & 0.597 \\ 0.35 & 0.65 & 0 \\ 0.2982 & 0.3275 & 0.3473 \end{bmatrix}$$

Figure 9: An example of transition matrix M.

To use Markov chain in modeling the transition in spatial distribution, we can design the spatial markov matrix. This matrix extends the traditional $k \times k$ transition matrix into a $k \times k \times k$ tensor. Conditioning on different response category of the spatial lag in the initial period, there will be k different transition probability matrix. The class of the neighbors is summarized by the spatial lag $z_i^* = \sum_{j=1}^{N} w_{i,j} z_j$. The overall influence of spatial dependence would be reflected in the differences

Table 1: A Spatial Markov Matrix

Spatial Lag	t_0	$t_1 = a$	$t_1 = b$	$t_1 = c$
a	a	$m_{aa\|a}$	$m_{ab\|a}$	$m_{ac\|a}$
	b	$m_{ba\|a}$	$m_{bb\|a}$	$m_{bc\|a}$
	c	$m_{ca\|a}$	$m_{cb\|a}$	$m_{cc\|a}$
b	a	$m_{aa\|b}$	$m_{ab\|b}$	$m_{ac\|b}$
	b	$m_{ba\|b}$	$m_{bb\|b}$	$m_{bc\|b}$
	c	$m_{ca\|b}$	$m_{cb\|b}$	$m_{cc\|b}$
c	a	$m_{aa\|c}$	$m_{ab\|c}$	$m_{ac\|c}$
	b	$m_{ba\|c}$	$m_{bb\|c}$	$m_{bc\|c}$
	c	$m_{ca\|c}$	$m_{cb\|c}$	$m_{cc\|c}$
not considered	a	m_{aa}	m_{ab}	m_{ac}
	b	m_{ba}	m_{bb}	m_{bc}
	c	m_{ca}	m_{cb}	m_{cc}

between the marginal cell values and the corresponding values in the various conditional matrices.

For example, if $m_{bc} > m_{bc|a}$ then the probability of an upward move for median class regions, irrespective of their neighbors, is higher than the probability of an upward move for median class regions with poor neighbors.

Multivariate Time Series

Another model for spatial-temporal lattice data is multivariate time series. Let $\boldsymbol{Z}_t = [Z(\boldsymbol{s_1}, \boldsymbol{t}), \cdots, Z(\boldsymbol{s_n}, \boldsymbol{t})]$ denote the vector of n responses across all spatial locations, at time t. A multivariate ARMA(p, q) process is given by:

$$\boldsymbol{Z}_t - \boldsymbol{A}_1 \boldsymbol{Z}_{t-1} - \cdots - \boldsymbol{A}_p \boldsymbol{Z}_{t-p} = \epsilon_t + \boldsymbol{B}_1 \epsilon_{t-1} + \cdots + + \boldsymbol{B}_q \epsilon_{t-q}$$

where $\boldsymbol{A}_1, \cdots, \boldsymbol{A}_p$ and $\boldsymbol{B}_1, \cdots, \boldsymbol{B}_q$ are $n \times n$ matrices. ϵ_t is a white noise process such that $\text{Var}(\epsilon_t) = \Sigma$ and $\text{Cov}(\epsilon_t, \epsilon_\tau) = 0$ for $t \neq \tau$. This model can be considered when the number of spatial locations is relatively small.

12.2 Spatial-Temporal Kriging

Kriging in space-time is pretty much the same as Kriging in space, with an extra dimension t. Let $\mu(\boldsymbol{s}, t)$ denote the mean of $Z(\boldsymbol{s}, t)$, the predictor is now formed as:

$$Z^*(\boldsymbol{s}, t) = \sum_{\alpha=1}^{N(\boldsymbol{s},t)} \lambda_\alpha Z(\boldsymbol{s}_\alpha, t_\alpha).$$

Similar to spatial Kriging, the key in prediction is to model the covariance structure between any two observations. By definition, covariance between two spacetime variables is $\text{Cov}[Z(\boldsymbol{s}, u), Z(\boldsymbol{r}, v)] = \mathbb{E}\{[Z(\boldsymbol{s}, u) - \mu(\boldsymbol{s}, u)][Z(\boldsymbol{r}, v) - \mu(\boldsymbol{r}, v)]\}$. As before, we normally require the stationarity of the spatial-temporal process. For example, the second-order stationarity of spatio-temporal data requires that

1. $\mathbb{E}\{[Z(\boldsymbol{s}, t)]\} = \mu$.
2. $\text{Cov}[Z(\boldsymbol{s}, u), Z(\boldsymbol{r}, v)] = C(\boldsymbol{r} - \boldsymbol{s}, v - u)$.

Although the definition of second-order stationarity in the spacetime context is a straightforward analogue extension, there are some fundamental differences between the spatial and spatial-temporal settings. In particular, in the spatial setting, $C(\boldsymbol{h}) = C(-\boldsymbol{h})$, by definition of the covariance function. However, this is not the case in the spatio-temporal setting for the time dimension.

Covariance functions for which $C(\boldsymbol{h}, u) = C(\boldsymbol{h}, -u)$ ($C(-\boldsymbol{h}, u) = C(\boldsymbol{h}, u)$) holds, for all $\boldsymbol{h}$ and u, are called fully symmetric. Among the class of fully symmetric covariances, a covariance function is separable if $C(\boldsymbol{h}, u)/C(\boldsymbol{h}, 0) = C(\boldsymbol{0}, u)/C(\boldsymbol{0}, 0)$, for all $\boldsymbol{h}$ and u. If this condition holds, we see that the space time covariance can be factored (separated) into the product of a purely spatial covariance

and a purely temporal covariance. It is straightforward to show that a separable covariance must be fully symmetric, but full symmetry does not imply separability.

The following covariance is an example from separable covariance function

$$C(\boldsymbol{h}, u) = \sigma^2 \exp(-\nu_s \|\boldsymbol{h}\|) \exp(-\nu_t |u|)$$

with $\sigma^2 > 0, \nu_s > 0, \nu_t > 0$. Nonseparable functions can also find applications, taking similar form as the following example.

$$C(\boldsymbol{h}, u, \theta) = \frac{\sigma^2}{(|u|^{2\gamma} + 1)^{\tau}} \exp\left[\frac{-c\|\boldsymbol{h}\|^{2\gamma}}{(|u|^{2\gamma} + 1)^{\beta\gamma}}\right]$$

Here, τ determines the smoothness of the temporal correlation; $\gamma \in (0, 1]$ determines the smoothness of the spatial correlation; c determines the strength of the spatial correlation; $\beta \in (0, 1]$ determines the strength of space/time interaction. In this parameterization, $\gamma = 1$ corresponds to the Gaussian covariance function, while $\gamma = 1/2$ corresponds to the exponential covariance function. Smaller value of γ leads to less smoothness in the interpolation results.

13 References

- Sherman, M. (2011). Spatial statistics and spatio-temporal data: covariance functions and directional properties. John Wiley & Sons.
- Geographic Information Technology Training Alliance (GITTA) http://www.gitta.info/website/en/html/inde
- Rey, S. J. (2001). Spatial empirics for economic growth and convergence. Geographical analysis, 33(3), 195-214.
- LeSage, J., & Pace, R. K. (2009). Introduction to spatial econometrics. Chapman and Hall/CRC.