1장 예측분석 소개

Predictive Analytics

- **Predictive analytics란?**

  대량의 데이터로부터 1) **규칙이나 패턴을 찾아서** **향후** **2) 새로운 자료(미래)에 대해서 예측**하는 과정으로 **데이터마이닝** 기법이나 **시계열분석** 기법 등 다양한 영역의 기법을 활용

- **Predictive analytics 데이터 구조**

  데이터마이닝에서 데이터셋을 모형구축에 사용될 **훈련용(train)** 데이터셋과 예측력 평가에 사용될 **평가용(test)** 데이터셋으로 나누어 모형 평가

  시계열 자료분석에서는 모형구축을 위한 **in-sample** 데이터셋과 예측력 평가를 위한 **out-of-sample** 데이터셋으로 구분

Predictive Analytics

- **데이터마이닝**

  대량의 데이터로부터 규칙이나 패턴을 찾아내는 과정으로 통계학, 데이터베이스, 기계학습, 인공지능의 다양한 영역의 기법을 포함

- **데이터마이닝 적용분야**

  일반기업: 표적마케팅, 고객세분화, 고객성향분석

  금융분야: 신용평가, 거래사기 적발

  제조업분야: 품질관리

  의학분야: 유전자 분석, 지구과학 및 천문분야의 자료처리

  빅데이터 분석분야: 텍스트마이닝, 음성 및 영상 분석

- **시계열자료분석**

  데이터가 **시계열의 특성(시간 주기)**을 가질 경우에는 시계열 자료분석 기법을 적용하여 시간상 **미래 자료**에 대한 **예측능력**을 향상시킬 수 있다.

- **시계열 분석 적용분야**

  전기자동차 차월 판매예측, 내일의 전기수요량 예측, 내년 여름 에어컨 판매량 예측

Predictive Analytics

- **지도학습(supervised learning): 예측모형**
  - 예측모형은 결과값이 알려진 다변량 자료를 이용하여 모형을 구축하고 이를 통해 새로운 자료에 대한 결과값에 대한 예측 및 분류가 주목적임
  - 결과값이 연속형인 경우에는 예측(prediction)이 목적이고, 이산형인 경우에는 분류(classification)가 주목적임
  - 로지스틱 회귀, 의사결정나무, 판별분석, 인접이웃분류, 베이즈 분류, 신경망, 서포트벡터머신, 앙상블

- **비지도학습(unsupervised learning)**
  - 예측모형과 달리 결과값을 요구하지 않는 자료에 대한 분석
  - 데이터 개체들 간의 유사성에 기반을 두고 전체 개체를 여러 개의 그룹으로 나눔
  - 군집분석(k-평균군집, 계층적 군집, 혼합분포군집), 주성분분석 등

데이터마이닝 알고리듬

- **Top 10 데이터마이닝 알고리듬**

  1) C4.5 (의사결정나무) 2) k-평균군집 3) SVM

  4) 연관분석 5) EM 6) 페이지 랭크

  7) 앙상블 8) K-NN 9) 단순베이즈 10) CART (의사결정나무)

- **Top 10 기계학습 알고리듬**

  A) **지도학습**:

  1) 의사결정나무 2) 단순베이즈 3) OLS 회귀 4) 로지스틱 회귀 5) SVM 6) 앙상블

  B) **비지도학습**:

  1) 군집분석 2) 주성분분석 3) 특잇값분석 4) 독립성분석

모형평가 방법

1) 최적 모형 선택 기준

- 결정계수: $R^2 = SSR/SST = 1-SSE/SST$
- 수정결정계수: $adj\_R^2 = 1-\{(n-1)/(n-p)\}\{SSE/SST\}$
- 평균제곱오차: $MSE = SSE/(n-p)$
- Mallow's C_p: $C_p = p+(MSE-\hat{\sigma}^2)(n-p)/\ \hat{\sigma}^2 = SSE/\hat{\sigma}^2+2p-n$

  ⇨ C_p의 값이 p와 가장 가까운 값을 가지는 p를 선택한다.

P: 모수의 수 (절편 포함)

SSE: 모형의 오차제곱합 $\sum(y_i-\hat{y}_i)^2$

$\hat{\sigma}^2$: 모든 예측변수를 포함한 적합모형의 평균제곱오차

모형평가 방법

2) 정보 기준 (information criteria)과 예측 제곱합

a) Information criteria

AIC = n ln(SSE/n) + 2p

BIC = n ln(SSE/n) + p ln(n)

APC = (n+p)/{n(n-p)} SSE (Amemiya 예측기준)

⇨ 이 들 값이 작을수록 정확함

b) 예측제곱합 (PRESS: prediction sum of square)

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2$$

여기서 $\hat{y}_{i(i)}$은 i 번째 자료를 제외하고 적합한 모형으로부터 i 번째 값을 추정한 것

c) 예측결정계수 (Predicted R^2)

pred_R^2 = 1 – PRESS / SST

모형평가 방법

3) 이진 반응변수의 경우

정오분류표		예측집단	
		C_1 (Y=1): 참	C_2 (Y=0): 거짓
실제집단	C_1 (Y=1): 참	f_{11} (true positive)	f_{12} (false negative)
	C_2 (Y=0): 거짓	f_{21} (false positive)	f_{22} (true negative)

① 정확도(accuracy) 또는 정분류율(correct classification rate): 전체에서 정확히 예측한 비율

$(f_{11}+f_{22})/n \Rightarrow$ 1에 가까울수록 바람직함

② 민감도(sensitivity): 실제 참인 것을 참으로 예측(분류)한 비율 (참긍정)

$f_{11}/(f_{11}+f_{12}) \Rightarrow$ 1에 가까울수록 바람직함

③ 특이도(specificity): 실제 거짓인 것을 거짓으로 제대로 예측(분류)한 비율 (참부정)

$f_{22}/(f_{21}+f_{22}) \Rightarrow$ 1에 가까울수록 바람직함

<참고> 정분류율 = $(f_{11}+f_{12})/n$ X 민감도 + $(f_{21}+f_{22})/n$ X 특이도

오차비율(error rate) = 1 – 정확도

모형평가 방법

3) 이진 반응변수의 경우 (계속)

정오분류표		예측집단	
		C_1 (Y=1): 참	C_2 (Y=0): 거짓
실제집단	C_1 (Y=1): 참	f_{11} (true positive)	f_{12} (false negative)
	C_2 (Y=0): 거짓	f_{21} (false positive)	f_{22} (true negative)

④ 정밀도(precision): 긍정예측(분류) 가운데 참긍정의 비율

$f_{11}/ (f_{11}+f_{21})$

⑤ 재현율(recall): 민감도와 유사. 높은 재현율은 높은 식별능력을 의미

$f_{11}/(f_{11}+f_{12})$

<참고> F 점수: 정밀도와 재현율을 조합하여 모델의 성능을 측정 (0과 1 사이의 값)

F 점수 = 2x정밀도x재현율/(정밀도+재현율)=$2xf_{11}/(2xf_{11}+f_{21}+f_{12})$

모형평가 방법

4) 연속 반응변수의 경우

① 평균절대오차

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

② 평균제곱오차

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

③ 평균절대백분위오차

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}\times 100$$

모형평가 방법

5) 모형선택을 위한 비교 방법

a) 신뢰구간 이용법(연속 반응변수의 경우 주로 사용)

K-중첩 교차타당법을 수행한 후 평균오차율을 추정하고 이를 이용하여 오차율 차이에 대한 신뢰구간을 구하여 비교하는 방법

b) ROC (Receiver Operating Characteristic) 곡선 이용법(이산 반응변수(0, 1)의 경우 주로 사용)

- 검증용 자료에 대해 예측값(주로 연속형 변수)을 내림차순으로 정렬한 뒤, 분류를 위해 기준값(0~1사이 값)을 선택하면 정오분류표를 얻게 된다.
- ROC 곡선은 기준값을 0에서 1 사이의 값으로 변환시키면서 해당 정오분류표로부터 거짓긍정(1-특이도)와 참긍정(민감도)의 값을 구하고, 이 값을 X, Y 좌표상에 연결해서 그린 그래프. 곡선 아래의 면적을 c 통계량 또는 AUC(area under curve)라고 하고 이 면적이 클수록 모형의 성능이 우수함을 나타낸다.

참고) 예측값이 기준값 보다 크면 1, 작으면 0으로 분류하여 실제 값과 분류값을 이용하여 정오분류표 작성

모형평가 방법

❖ 교차타당법(cross validation method)

- 데이터마이닝에서 데이터셋을 모형구축에 사용될 **훈련용(train)** 셋과 예측력 평가에 사용될 **평가용(test)** 셋으로 나누어 모형 평가
- 데이터 양이 충분히 많으면, 예측용: 평가용을 50:50으로 랜덤하게 나누어 적용

a) K-중첩(fold) 법

① 데이터 양이 충분치 않을 경우에 전체 데이터셋을 K 조각으로 나누고, K-1 조각으로 모형 구축한 뒤 나머지 조각에 대해 예측 수행.

② 이러한 절차를 K 번 반복 수행하여 평가

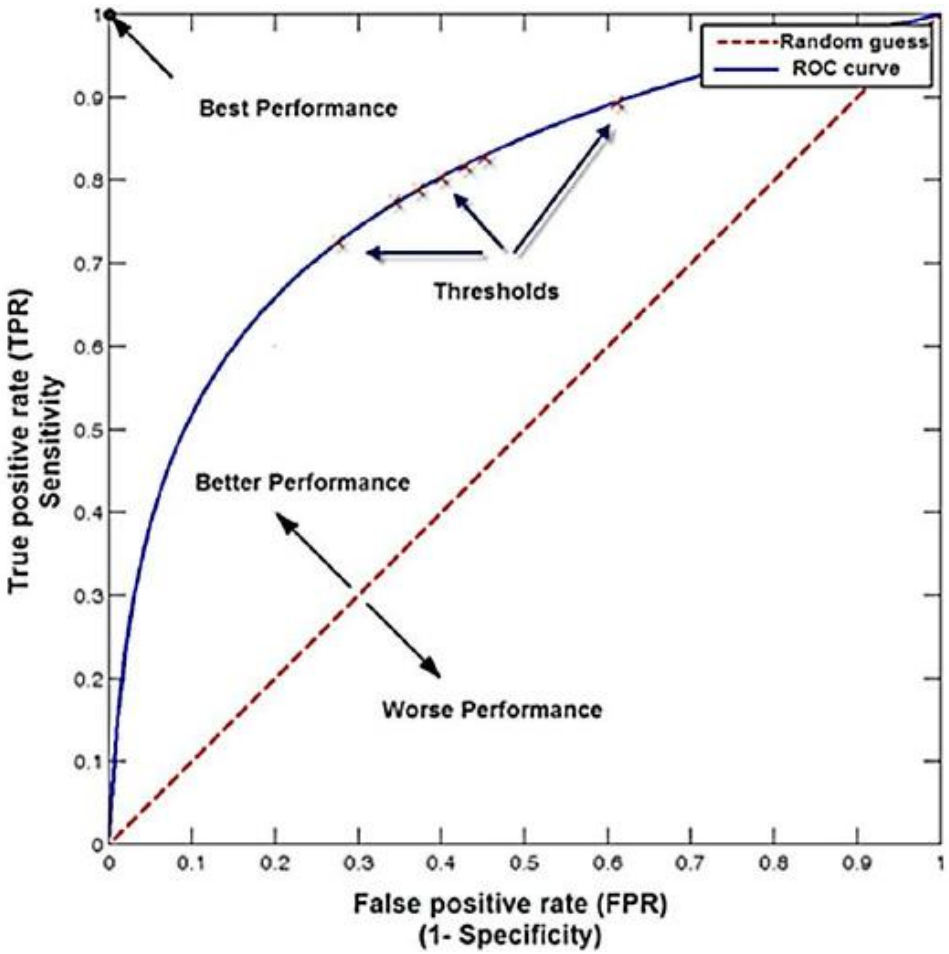
b) Leave one out 방법

① K = n 이라고 생각하고, 진행. 즉, 한 개의 데이터만 빼고 나머지로 모형 구축한 뒤 나머지 한 개에 대해서 예측 실행

② 이러한 절차를 n 번 반복 수행하여 평가

모형평가 방법

❖ **ROC 곡선**



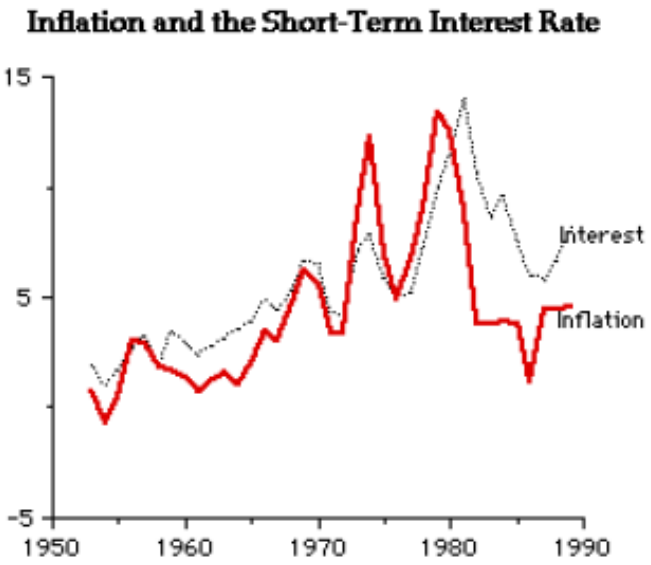
- AUC 점수 체계(rule of thumb)
  - 0.9~1.0: 탁월하다
  - 0.8~0.9: 뛰어나다
  - 0.7~0.8: 괜찮다
  - 0.6~0.7: 형편없다
  - 0.5~0.6: 가치없다

시계열자료 (Time Series Data)

- A *time series variable* (Y) consists of data observed over n periods of time.
- Businesses use time series data
  - to monitor a process to determine if it is stable
  - to predict the future (forecasting)
- Time series data are usually plotted as a line graph.
- Time is on the horizontal (X) axis.
- This reveals how a variable changes over time.
- Fluctuations are easier to see on a line graph.

시계열자료

- Example:



- The following notation is usually used:

  y_t is the value of the time series in period t

  t is an index denoting the time period $(t = 1, 2, \ldots, n)$

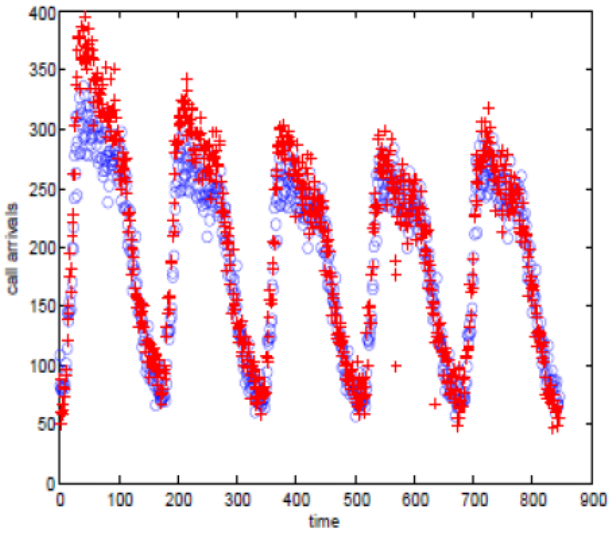
  n is the number of time periods

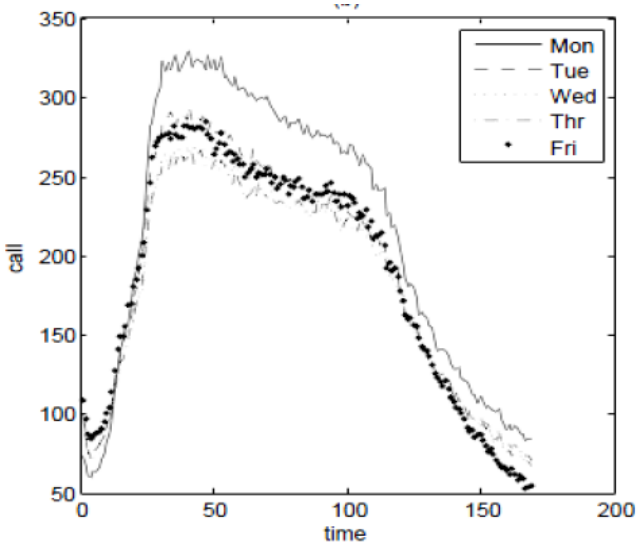
  $y_1, y_2, \ldots, y_n$ are the data set for analysis

시계열자료 측정

- Time series data may be measured *at a point in time*.
- For example, *prime rate of interest* is measured at a particular point in time.
- Time series data may also be measured *over an interval of time*.
- For example, *Gross Domestic Product* (*GDP*) is a flow of goods and services measured over an interval of time.
- Data can be collected once every decade/ year (e.g., 1 observation per year)/ quarter (e.g., 4 observations per year)/ month (e.g., 12 observations per year)/ week/ day/ hour
- Same time series data can be plotted in a different way.

시계열자료 측정





- Types of variables on y-axis (observation value) are usually discrete (ex: call arrivals per minute at a call center) or continuous (ex: monthly height of kids).
- Numerical variable has two types: Discrete and Continuous

  Discrete: countable, possible values are already decided

  Ex) call arrivals (number of phone calls),…

  Continuous: non-countable, any value within an interval

  Ex) stock price, height, GNP,…

시계열 모델링 기법

- Decomposition model
- Autoregressive (AR) model
- Autoregressive Integrated Moving Average (ARIMA) model
- ARMAX model
- Spectral Analysis
- Generalized Additive Model (GAM)
- Panel Data Analysis
- Extreme Value Analysis

시계열 모델링 기법

- Time series decomposition: additive or multiplicative models (regression model)

  Ex) $y_t = \alpha_0 + \alpha_1 x_{1t} + \cdots + \alpha_p x_{pt} + \varepsilon$

- ARIMA and SARIMA models: recursive model (autoregressive model)

  Ex) $y_t = \alpha_0 + \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + \varepsilon$

- ARIMAX models: ARIMA with exogeneous variable

  Ex) $y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 x_{t-1} + \varepsilon$

- GARCH models: recursive model

  Ex) $r_t = \varepsilon_t \sqrt{\alpha_0 + \alpha_1 r_{t-1}^2}$

- Spectral analysis: frequency domain approach ▷

  find significant periodicity (using periodogram)

시계열 모델링 기법

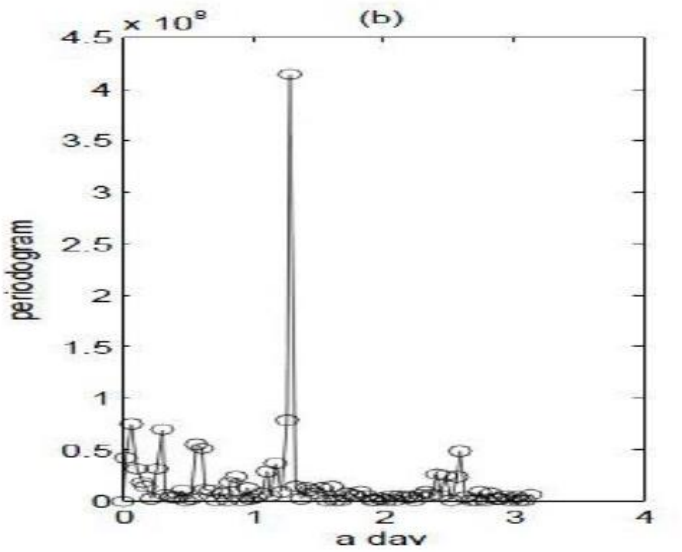
- Generalized additive model

  Ex) $y_t = \alpha_0 + f_1(x_{t-1}) + f_2(x_{t-2}) + \varepsilon$

- Extreme value model (block/ threshold approach): generalized extreme value distribution model, generalized Pareto distribution model

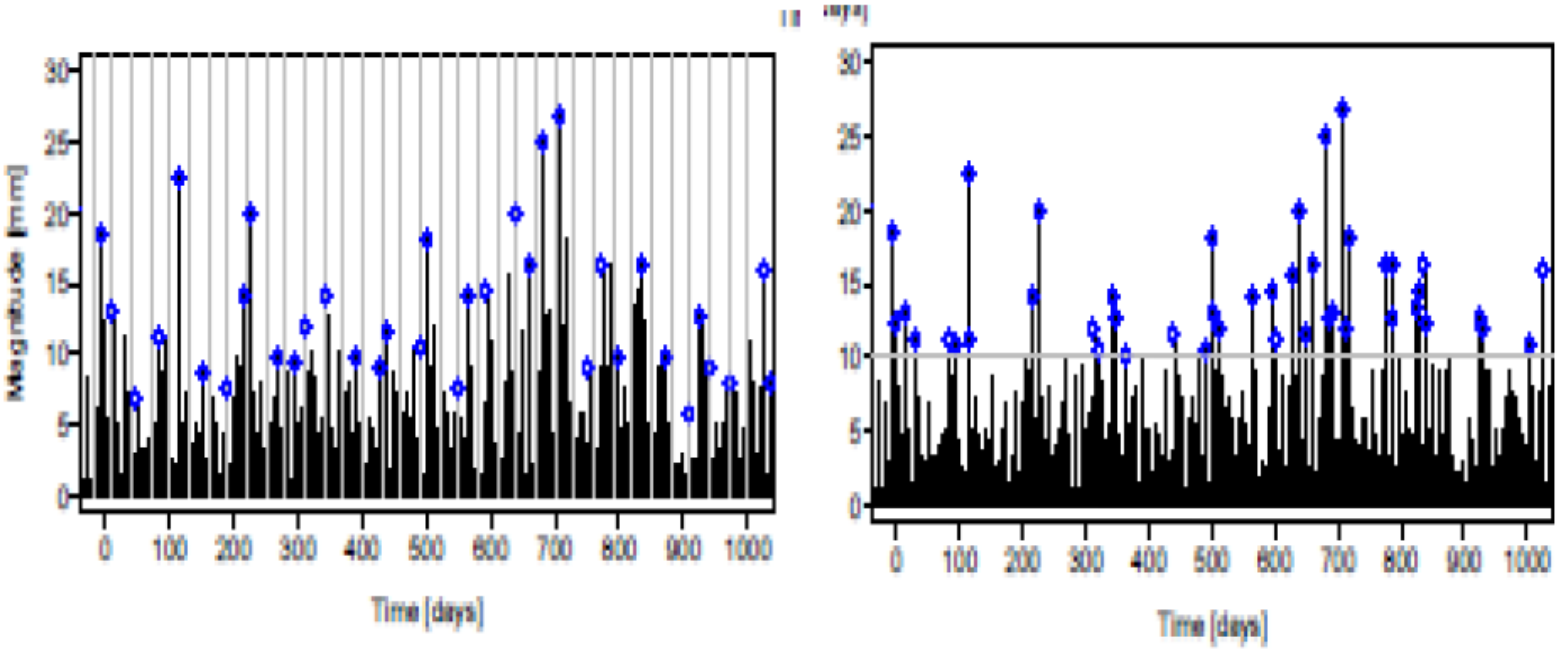
- Longitudinal data analysis (repeated measure analysis)

- Periodogram plot



시계열 모델링 기법

- Block maxima and peak over threshold plots

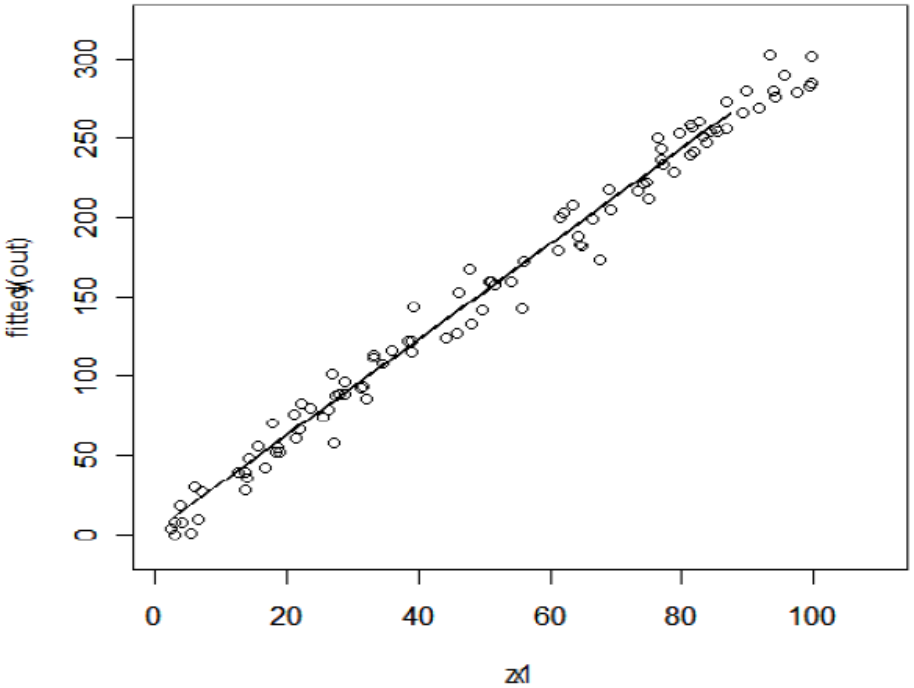


Stationary Processes

- often a time series has same type of random behavior from one time period to the next
  - outside temperature: each summer is similar to the past summers
  - interest rates and returns on equities
- *stationary stochastic processes* are probability models for such series
- a process is stationary if its behavior is unchanged by shifts in time
- a process is ***weakly stationary*** if its mean, variance, and covariance are unchanged by time shifts
- in the real world, many financial time series y_t are not stationary but the *changes* (or *first order difference:* $z_t = y_t - y_{t-1}$) in these time series may be stationary
- Lag operator B: $(1-B)X_t = X_t - X_{t-1}$
- a process should be stationary in order to apply some forecasting models (ex: ARMA)
- a process does not have to be stationary to some forecasting models (ex: ARIMA)

모델추정(in-sample)과 예측정확성 평가(out-of-sample)

- The ultimate goal is forecasting the unobserved future data point using some good forecasting model.
- We need to select a model by evaluation its performance.
- Model performance is measured using the observed historical past data.
- Example: $y = \beta_0 + \beta_1 x + e$, $e \sim N(0, \sigma^2)$: model, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$: estimated line



모델성능(예측정확성)평가와 *k*-time ahead prediction

- "Fit" refers to how well the estimated or predicted data using forecasting models matches the observed historical past data.
- The observed historical past data used for the estimation of the forecasting model are referred to as the in-sample data.
- The observed historical past data used for the prediction via the estimated forecasting model are referred to as the out-of-sample data.
- In-sample fit (or accuracy) & out-of-sample fit can be considered.
- In-sample fit is related with the estimated data and the historical data.
- Out-of-sample fit is related with the predicted data and historical data.
- One time ahead forecast (can be treated as short-term forecasting) or *k* time ahead forecast (can be treated as long-term forecasting) can be considered.