데이터과학을 위한 R프로그래밍

7주차. 상관분석과 회귀모형

이혜선 교수
포항공과대학교 산업경영공학과

목차

7주차. 상관분석과 회귀모형

7주차

3차시

회귀분석의 진단과 평가

POSTECH

회귀분석 - 데이터

SF 데이터 (항공 출도착 지연 데이터)

1. year: multi-valued discrete (출발연도 : 정수값)
2. month: multi-valued discrete (출발월 : 정수값)
3. day: multi-valued discrete (출발일 : 정수값)
4. dep_time: multi-valued discrete (실제출발시간 : 정수값)
5. sched_dep_time: multi-valued discrete (출발시간 : 정수값)
6. dep_delay: multi-valued discrete (출발지연시간 : 정수값)
7. arr_time: multi-valued discrete (실제도착시간 : 정수값)
8. sched_arr_time: multi-valued discrete (도착시간: 정수값)
9. arr_delay: multi-valued discrete (도착지연시간 : 정수값)
10. carrier: string (unique for each instance) (항공사 이름)
11. flight: string (unique for each instance) (항공기 번호)
12. tailnum: string (unique for each instance) (항공기 고유번호)
13. origin: string (unique for each instance) (출발공항 이름)
14. dest: string (unique for each instance) (도착공항 이름)
15. air_time: multi-valued discrete (비행시간 : 정수값)
16. distance: continuous (비행거리 : 연속형변수)
17. hour: multi-valued discrete (출발시간 : 정수값)
18. minute: multi-valued discrete (출발분 : 정수값)
19. time_hour: multi-valued discrete (출발 시각 : yy-mm-dd hh-mm-ss)

```
# Regression with flight(NY->SF)

library(dplyr)
library(ggplot2)

# set working directory
setwd("D:/tempstore/moocr")

# subset of flight data in SFO (n=2974)
# dest="SFO", origin=="JFK", arr_delay<420, arr_delay>0
SF<-read.csv("SF_2974.csv")
head(SF)
str(SF)
```

```
> SF<-read.csv("SF_2974.csv")
> head(SF)
  year month day dep_time sched_dep_time dep_delay arr_time sched_arr_time arr_delay
1 2013     1   1      611            600        11      945            931        14
2 2013     1   1      745            745         0     1135           1125        10
3 2013     1   1     1029           1030        -1     1427           1355        32
4 2013     1   1     1112           1100        12     1440           1438         2
5 2013     1   1     1124           1100        24     1435           1431         4
6 2013     1   1     1436           1435         1     1840           1820        20
  carrier flight tailnum origin dest air_time distance hour minute           time_hour
1      UA    303  N532UA    JFK  SFO      366     2586    6      0 2013-01-01 06:00:00
2      AA     59  N336AA    JFK  SFO      378     2586    7     45 2013-01-01 07:00:00
3      AA    179  N325AA    JFK  SFO      389     2586   10     30 2013-01-01 10:00:00
4      UA    285  N517UA    JFK  SFO      364     2586   11      0 2013-01-01 11:00:00
5      B6    641  N590JB    JFK  SFO      349     2586   11      0 2013-01-01 11:00:00
6      DL   1322  N722TW    JFK  SFO      375     2586   14     35 2013-01-01 14:00:00
```

POSTECH

데이터의 기술통계치

- 기술통계치 : summary 함수 사용 (최소값, 중위수, 평균, 최대값 등)

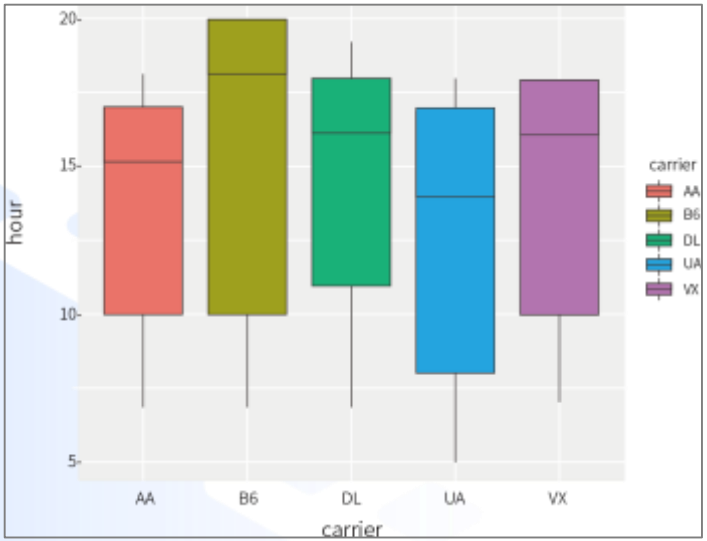
```
> summary(SF)
      year          month             day           dep_time    sched_dep_time   dep_delay
 Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 559   Min.   :-15.00
 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:1026   1st Qu.:1019   1st Qu.: -2.00
 Median :2013   Median : 7.000   Median :16.00   Median :1551   Median :1530   Median :  8.00
 Mean   :2013   Mean   : 6.903   Mean   :15.75   Mean   :1438   Mean   :1392   Mean   : 33.13
 3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1838   3rd Qu.:1755   3rd Qu.: 47.00
 Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2356   Max.   :2029   Max.   :396.00

    arr_time    sched_arr_time   arr_delay       carrier        flight          tailnum      origin
 Min.   :   1   Min.   :   3   Min.   :  1.00   AA:614   Min.   :  11.0   N502UA :  80   JFK:2974
 1st Qu.:1218   1st Qu.:1334   1st Qu.:  8.00   B6:404   1st Qu.:  85.0   N505UA :  69
 Median :1758   Median :1858   Median : 21.00   DL:534   Median : 303.0   N557UA :  64
 Mean   :1596   Mean   :1729   Mean   : 42.48   UA:932   Mean   : 427.2   N560UA :  64
 3rd Qu.:2128   3rd Qu.:2120   3rd Qu.: 53.00   VX:490   3rd Qu.: 595.0   N508UA :  62
 Max.   :2400   Max.   :2359   Max.   :411.00            Max.   :2915.0   N525UA :  61
                                                                          (Other):2574
  dest          air_time        distance         hour           minute                    time_hour
 SFO:2974   Min.   :304.0   Min.   :2586   Min.   : 5.00   Min.   : 0.00   2013-01-05 07:00:00:   5
            1st Qu.:344.0   1st Qu.:2586   1st Qu.:10.00   1st Qu.: 0.00   2013-03-08 07:00:00:   5
            Median :356.0   Median :2586   Median :15.00   Median :30.00   2013-01-09 07:00:00:   4
            Mean   :355.5   Mean   :2586   Mean   :13.67   Mean   :24.47   2013-01-16 07:00:00:   4
            3rd Qu.:367.0   3rd Qu.:2586   3rd Qu.:17.00   3rd Qu.:35.00   2013-02-04 10:00:00:   4
            Max.   :490.0   Max.   :2586   Max.   :20.00   Max.   :59.00   2013-02-08 07:00:00:   4
                                                                           (Other)            :2948
```

POSTECH

데이터 시각화

데이터 시각화 : 항공사별 출발시간

```
# Visualization : boxplot
ggplot(SF, aes(x=carrier, y=hour)) + geom_boxplot(aes(fill=carrier))
# boxplot(hour~carrier, data=SF, col=c("coral", "green", "orange","yellow",
```

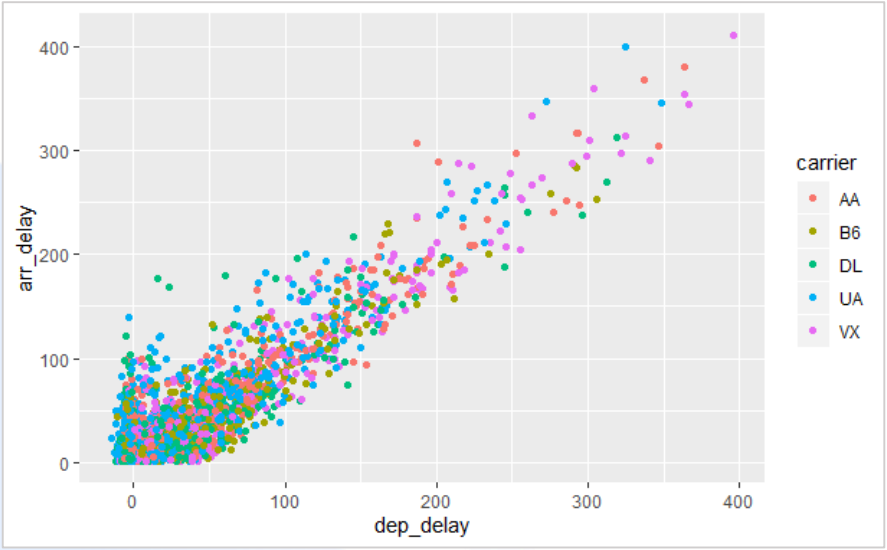


POSTECH

데이터 시각화

데이터 시각화 : 상관관계 (출발지연시간, 도착지연시간)

```
# Visualization : scatterplot
ggplot(SF, aes(arr_delay, dep_delay,color=carrier)) + geom_point( )
```



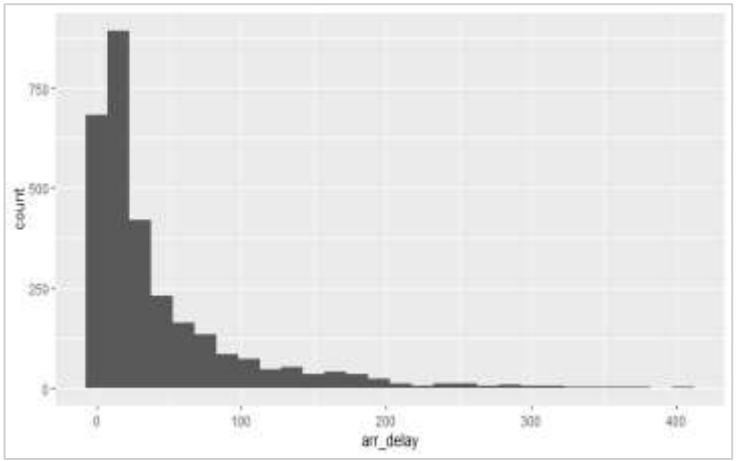
POSTECH

데이터 시각화

☑ Histogram
: ggplot(data, aes(x변수)) + geom_histogram()

```
# Visualization using dplyr : Histogram
SF %>%
  ggplot(aes(arr_delay)) + geom_histogram(binwidth = 15)
```

Histogram
: arr_delay(도착지연시간)



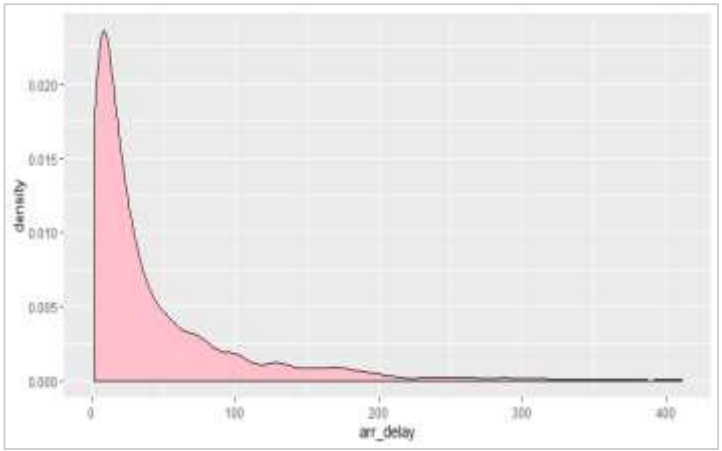
POSTECH

데이터 시각화

☑ 분포함수 추정 (Density Function)
: ggplot(data, aes(x변수)) + geom_density()

```
# Visualization using dplyr : Density Graph
SF %>%
  ggplot(aes(arr_delay)) + geom_density(fill = "pink",binwidth = 15)
```

Density Function
: arr_delay(도착지연시간)



회귀분석의 진단과 평가

POSTECH

데이터 시각화

Box Plot : ggplot(data, aes(x변수,y변수)) +geom_boxplot()

2. 상자그림

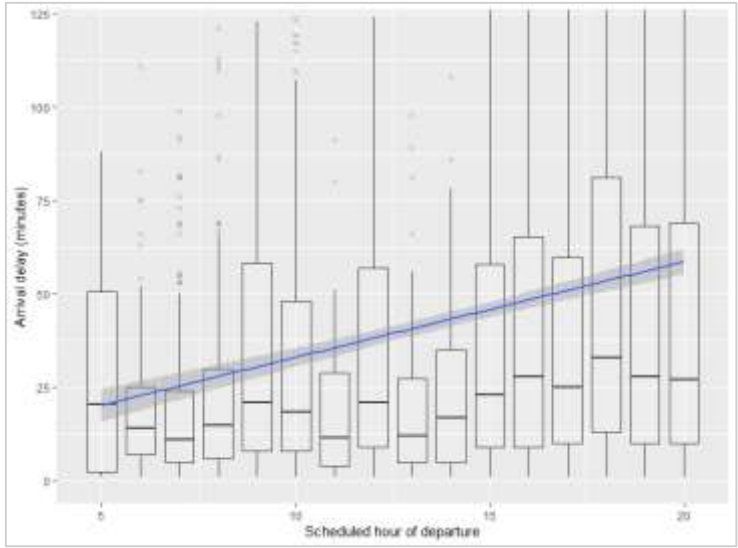
x변수 : hour(출발시각) y변수 : arr_delay(도착지연시간)

```
# Visualization using dplyr : Box-Plot by departing time
SF %>%
  ggplot(aes(x = hour, y = arr_delay)) +
  geom_boxplot(alpha = 0.1, aes(group = hour)) + geom_smooth(method = "lm") +
  xlab("Scheduled hour of departure") + ylab("Arrival delay (minutes)") +
  coord_cartesian(ylim = c(0, 120))
```

geom_smooth(method = “lm”)을 통해 Linear Graph를 Box Plot 위에 그림

X축과 y축 이름을 각각 "Scheduled hour of departure“, "Arrival delay (minutes)"로 설정

Y축의 범위는 0부터 120으로 설정



POSTECH

회귀분석 - 단순회귀모형

단순회귀모형 : lm(y변수~x변수, data=)

단순회귀모형

종속변수 : arr_delay(도착지연시간), 독립변수: hour(출발시간)

```
# linear regression
m1<- lm(arr_delay ~ hour , data = SF)
summary(m1)
anova(m1)
```

```
> m1<- lm(arr_delay ~ hour , data = SF)
> summary(m1)

Call:
lm(formula = arr_delay ~ hour, data = SF)

Residuals:
   Min     1Q Median     3Q    Max
-57.64 -32.09 -17.98  10.91 362.58

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.5448     3.3101   2.279   0.0227 *
hour          2.5549     0.2306  11.077   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.79 on 2972 degrees of freedom
Multiple R-squared:  0.03965,   Adjusted R-squared:  0.03932
F-statistic: 122.7 on 1 and 2972 DF,  p-value: < 2.2e-16
```

선형회귀식

y(arr_delay)=7.54+2.55(hour)

선형회귀식의 결정계수

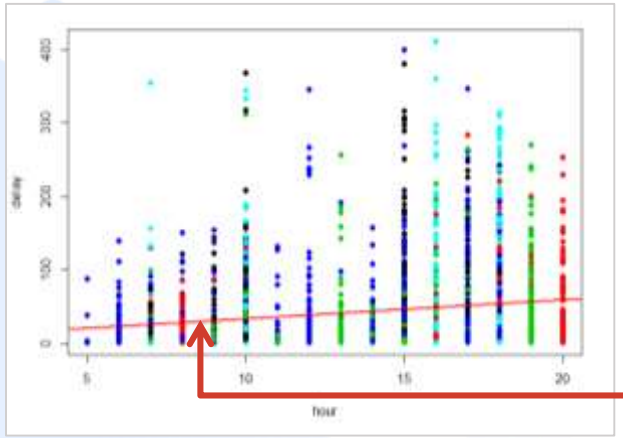
R^2=0.03965

POSTECH

회귀분석 - 단순회귀모형

산점도에 회귀선 그리기

```
# scatterplot with best fit lines
par(mfrow=c(1,1))
plot(SF$hour,SF$arr_delay , col=as.integer(SF$carrier), pch=19, xlab="hour",ylab="delay")
# best fit linear line
abline(lm(SF$arr_delay~SF$hour), col="red", lwd=2, lty=1)
```



plot(x축변수, y축변수)

abline : add line (선을 추가하는 함수)

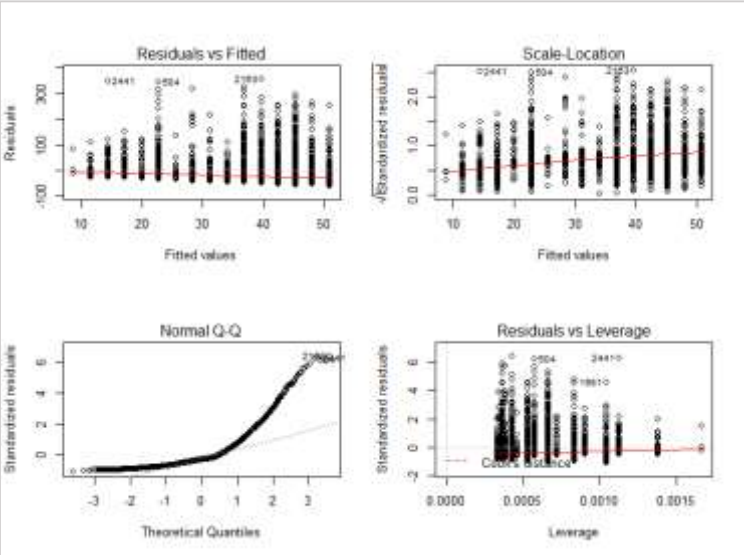
lm(y변수~x변수) : lm은 linear model(선형모형)의 약자

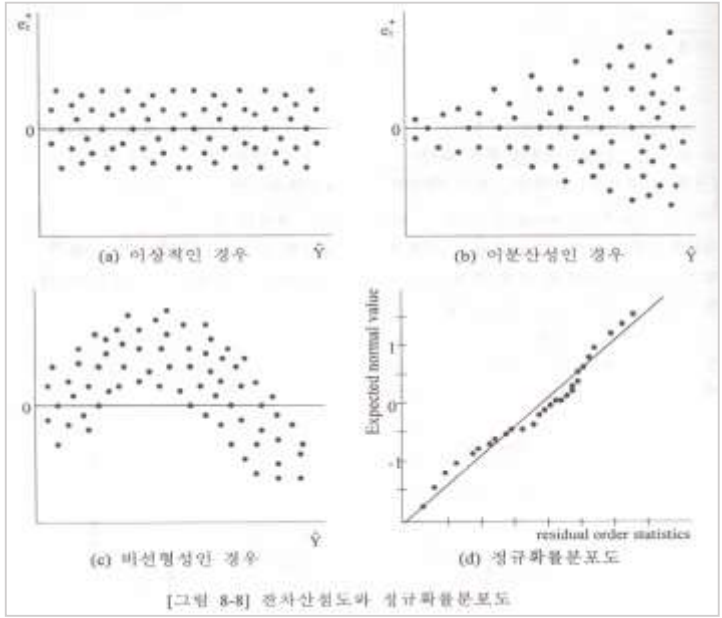
y(arr_delay)=7.54+2.55 (hour)

회귀분석 – 잔차의 산점도

회귀분석의 가정과 진단

```
# residual diagnostic plot
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
plot(m1)
```





[그림 8-8] 잔차산점도와 정규확률분포도

출처 전치혁, 정민근, 이혜선, 공학응용통계, 홍릉출판사, 2012