- 연습문제 1

1) 자료 파일: p2-1.dat

```
> str(p21)
'data.frame':    130 obs. of  7 variables:
 $ V1: int  1 2 3 4 5 6 7 8 9 10 ...
 $ V2: num  24946 11474 26131 44608 6262 ...
 $ V3: num  0.188 0.072 0.096 0.105 0.082 0.174 0.049 0.104 0.23 0.142 ...
 $ V4: num  0.068 0.038 0.085 0.129 0.111 0.074 0.138 0.076 0.049 0.081 ...
 $ V5: num  0.33 0.54 0.12 0.18 0.09 0.33 0.2 0.1 0.21 0.12 ...
 $ V6: int  1 1 1 1 1 1 1 1 2 2 ...
 $ V7: int  1 1 1 1 1 1 1 1 1 1 ...
```

V1: ID, V2: 총자산규모, V3: 레버리지 척도, V4: 수익성 척도
V5: 불안정성 척도 V6: 주식 등급
V7: 반응변수(1=투자성 등급, 0=투기성 등급)

2) 분석 내용

- 자료의 20%를 임의로 추출하여 test data로 분류:

```
> p21 <- read.table("D:/data/p2-1.dat")
> names(p21) <- c("id", paste("x", 1:5, sep=""), "y")

> set.seed(1234)

> x.id <- sample(1:nrow(p21), nrow(p21)*.2)

> p21_t <- p21[x.id,]

> p21_d <- p21[-x.id,]
```

- 나머지 80% training data로 최적 모형 구축
  - 모형 가정 확인
  - 이상값 탐지

- 최적 모형으로 test data를 대상으로 정분류율, AUC 계산

- 모형 구축

  1) 전진선택법

```
Single term additions

Model:
y ~ x5 + x2
       Df Deviance    AIC     LRT Pr(Chi)
<none>      52.176 58.176
x1      1   50.131 58.131 2.04500  0.1527
x3      1   52.151 60.151 0.02491  0.8746
x4      1   50.039 58.039 2.13666  0.1438
```

전진선택법 결과: (X_2, X_5)

2) 후진소거법

```
Single term deletions

Model:
y ~ x2 + x5
       Df Deviance    AIC    LRT   Pr(Chi)
<none>      52.176 58.176
x2      1   59.647 63.647  7.471  0.006269 **
x5      1   95.728 99.728 43.552 4.128e-11 ***
```

후진소거법 결과: (X_2, X_5)

3) 단계별선택법

단계별선택법 결과: (X_2, X_5) 포함

4) 모든 가능한 회귀

```
BIC
BICq equivalent for q in (0.00659589364266, 0.64585640166)
Best Model:
               Estimate Std. Error   z value     Pr(>|z|)
(Intercept)  19.436049  4.2970981  4.523064 6.095086e-06
x2          -13.699423  4.2549499 -3.219644 1.283499e-03
x5           -3.478617  0.8569975 -4.059075 4.926747e-05
```

```
AIC
BICq equivalent for q in (0.78102409649, 0.85240202462067)
Best Model:
                  Estimate   Std. Error   z value     Pr(>|z|)
(Intercept)  2.122026e+01 4.9882946502  4.254010 2.099759e-05
x1           2.598379e-04 0.0001167972  2.224694 2.610181e-02
x2          -1.549807e+01 4.8538590615 -3.192938 1.408333e-03
x4          -1.173097e-01 0.0822840988 -1.425667 1.539644e-01
x5          -3.825987e+00 1.0040346543 -3.810612 1.386231e-04
```

비유의적 변수 포함

5) AIC에 의한 단계적선택법

```
Coefficients:
(Intercept)           x1           x2           x4           x5
 16.0706006    0.0001733   -9.6331740   -0.1277116   -3.0848381

Degrees of Freedom: 103 Total (i.e. Null);  99 Residual
Null Deviance:     136.5
Residual Deviance: 47.53      AIC: 57.53
```

6) BIC에 의한 단계적선택법

```
Coefficients:
(Intercept)           x2           x5
     15.613       -9.086       -2.990

Degrees of Freedom: 103 Total (i.e. Null);  101 Residual
Null Deviance:     136.5
Residual Deviance: 52.18      AIC: 58.18
```

- 잠정 모형

(1) 모형 M_1: (X_1, X_2, X_4, X_5) 포함

- AIC에 의한 모형 선택

(2) 모형 M_2: (X_2, X_5) 포함

- 검정에 의한 모형 선택
- BIC에 의한 모형 선택

- 잠정 모형의 적합도 비교

```
> fit_m1 <- glm(y ~ . -id -x3, family=binomial, p21_d)
> fit_m2 <- glm(y ~ x2 + x5, family=binomial, p21_d)
```

  - 정분류율
  - ROC curve and AUC
  - AIC
  - BIC

- 정분류율

```
> CCR <- function(d, y, pred){
    y_hat <- (pred>=d)*1
    my_table <- table(y,y_hat)
    ccr <- sum(diag(my_table))/sum(my_table)*100
    res <- c(d, ccr)
    return(res)
  }
```

```
> CCR(d=0.5, y=p21_d$y, pred=fit_m1$fitted)
[1]  0.50000 91.34615
> CCR(d=0.5, y=p21_d$y, pred=fit_m2$fitted)
[1]  0.50000 89.42308
```

정분류율 측면: 모형 M_1이 약간 우세(큰 차이 없음)

- ROC curve

```
> library(pROC)

> roc(p21_d$y, fit_m1$fitted, percent=TRUE)$auc
Area under the curve: 97.01%

> roc(p21_d$y, fit_m2$fitted, percent=TRUE)$auc
Area under the curve: 96.23%
```

AUC 측면: 모형 M_1이 약간 우세(큰 차이 없음)

- AIC and BIC

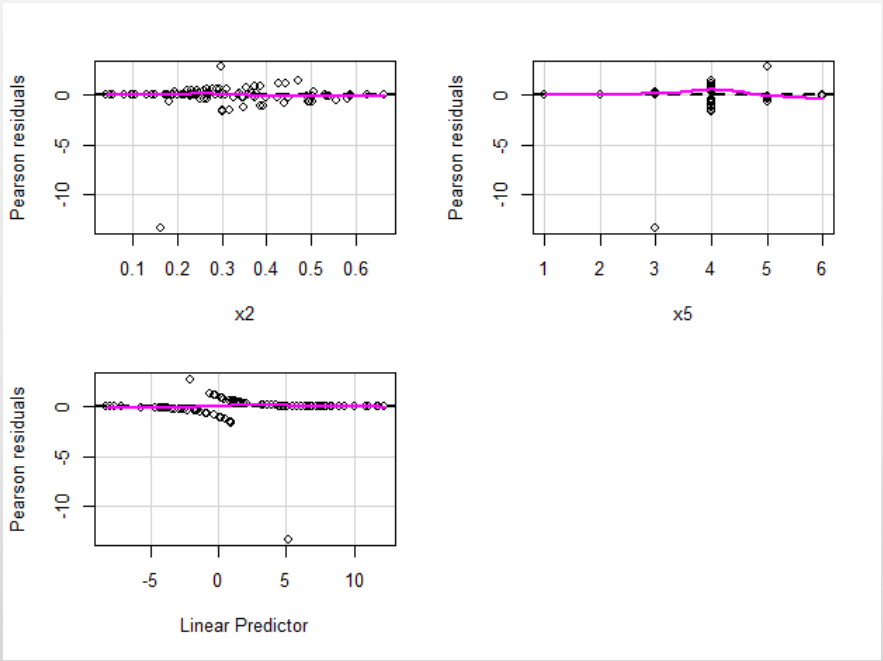
```
> AIC(fit_m1); AIC(fit_m2)
[1] 57.5306
[1] 58.17587

> BIC(fit_m1); BIC(fit_m2)
[1] 70.75256
[1] 66.10905
```

- 적합도에서 M_1과 M_2 사이에 큰 차이는 없는 것으로 보임
- 모수절약의 원칙에 따라 M_2를 선택
- 진단과정에서 모형이 변경될 가능성도 있음

- 모형 M_2의 진단

- 잔차 산점도

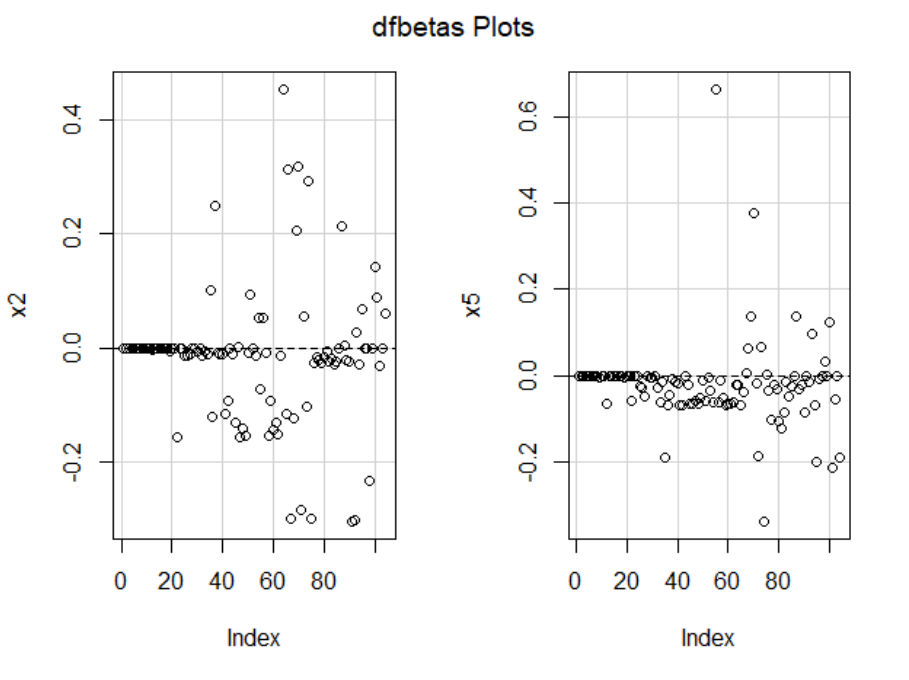
```
> library(car)
> residualPlots(fit_m2)
    Test stat Pr(>|Test stat|)
x2     2.4052             0.1209
x5     0.8337             0.3612
```



- 모형 설정에는 문제가 없어 보임
- 이상값 존재 가능성

- Dfbeta Plot

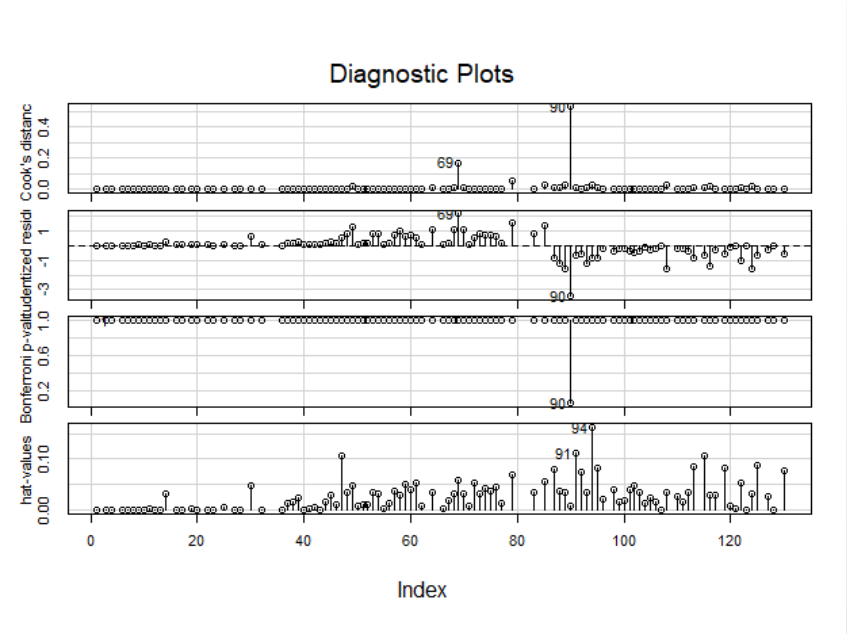
```
> dfbetasPlots(fit_m2)
```



영향력이 큰 관찰값
존재 가능성

- Influence Index Plot

```
> infIndexPlot(fit_m2)
```

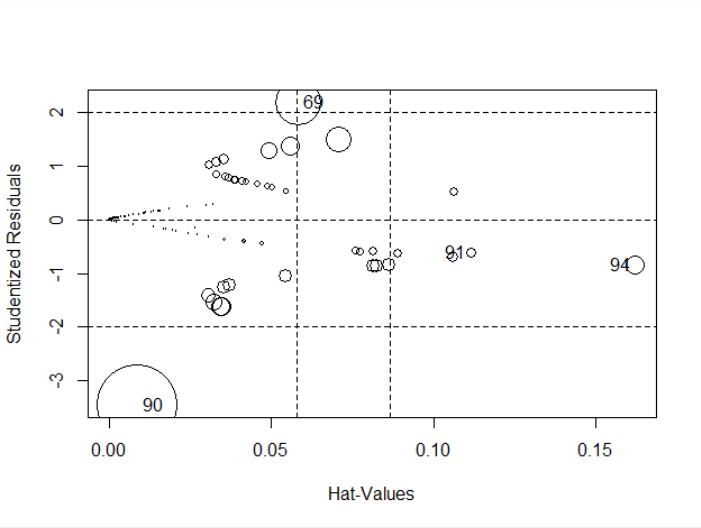


90번째 관찰값

- Cook's distance, Studentized residual, leverage 값이 모두 크며, Bonferroni p-value가 매우 작음

- Influence Plot

```
> influencePlot(fit_m2)
      StudRes          Hat       CookD
69  2.1974692 0.058212703 0.170433236
90 -3.4576265 0.008824531 0.532306099
91 -0.6145715 0.111444461 0.009095176
94 -0.8390101 0.162174871 0.028619932
```




```
> library(dplyr)

> p21_d %>% group_by(y) %>%
    summarize(m_2=mean(x2), m_5=mean(x5))
# A tibble: 2 x 3
      y   m_2   m_5
  <int> <dbl> <dbl>
1     0 0.415  4.79
2     1 0.249  2.85

> filter(p21_d, id==90)
  id   x1    x2    x3   x4 x5 y
1 90 3034 0.161 0.013 4.05  3 0
```

90번째 자료 삭제

- 90번째 관찰값 제거 후 재분석

  - 제거

```
> p21_d_1 <- filter(p21, id!=90)
```

  - 모형 선택

```
BIC
BICq equivalent for q in (0.0065958936426617 8,
0.645856401669784)
Best Model:
              Estimate Std. Error   z value     Pr(>|z|)
(Intercept)  19.436049  4.2970981  4.523064 6.095086e-06
x2          -13.699423  4.2549499 -3.219644 1.283499e-03
x5           -3.478617  0.8569975 -4.059075 4.926747e-05
```

동일한 모형 선택
- 하나의 관찰값만이 제거된 상태
- 모형 자체가 바뀌기는 어려운 상황

- 제거 전후 모형 비교

```
> fit_m2_1 <- glm(y ~ x2 + x5, family=binomial, p21_d_1)
> fit_m2 <- glm(y ~ x2 + x5, family=binomial, p21_d)
```

```
> library(car)
> compareCoefs(fit_m2, fit_m2_1)
Calls:
1: glm(formula = y ~ x2 + x5, family = binomial, data = p21_d)
2: glm(formula = y ~ x2 + x5, family = binomial, data = p21_d_1)

             Model 1 Model 2
(Intercept)   15.61   19.44
SE             3.68    4.30

x2            -9.09  -13.70
SE             3.66    4.25

x5           -2.990  -3.479
SE            0.781   0.857
```

- 회귀계수의 추정값에 큰 변화

- 적합도 비교

```
> AIC(fit_m2); AIC(fit_m2_1)
[1] 58.17587
[1] 56.89491
> BIC(fit_m2); BIC(fit_m2_1)
[1] 66.10905
[1] 65.47434
```

- AIC와 BIC 값: 감소

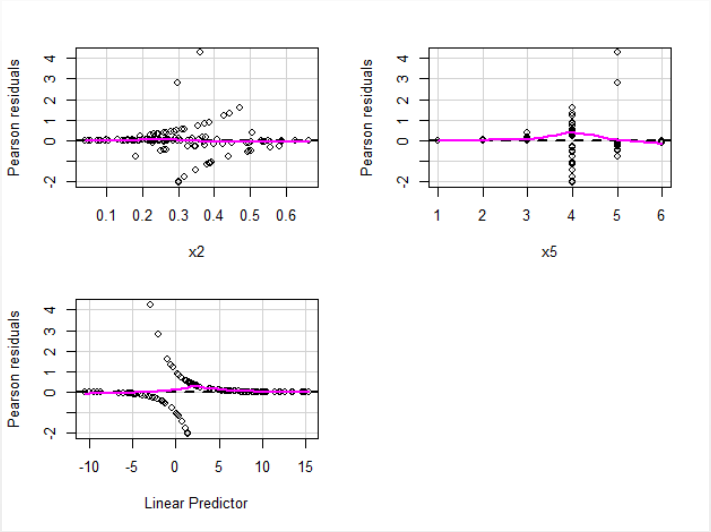
```
> roc(p21_d$y, fit_m2$fitted, percent=TRUE)$auc
Area under the curve: 96.23%

> roc(p21_d_1$y, fit_m2_1$fitted, percent=TRUE)$auc
Area under the curve: 97.57%
```

- AUC 값 약간의 상승

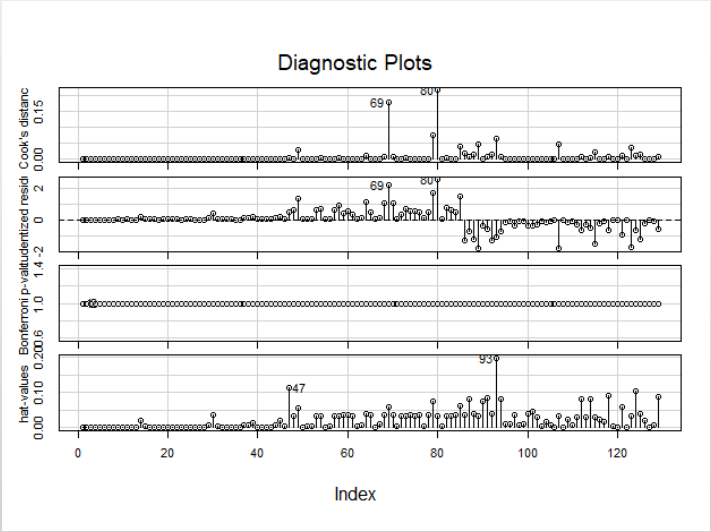
- 90번째 관찰값의 제거는 전체적으로 모형의 적합도를 향상 시킨 것으로 보임.

- 제거 후 모형에 대한 진단

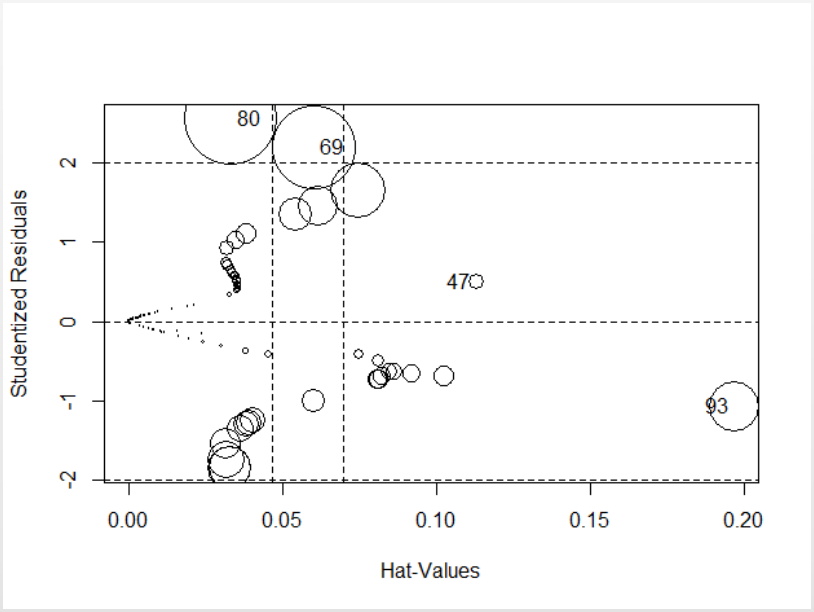
```
> residualPlots(fit_m2_1)
   Test stat Pr(>|Test stat|)
x2    0.1399             0.7084
x5    0.2555             0.6132
```




```
> infIndexPlot(fit_m2_1)
```



```
> influencePlot(fit_m2_1)
      StudRes         Hat        CookD
47  0.5074175 0.11298299 0.006133818
69  2.2012855 0.06005216 0.176554129
80  2.5569570 0.03322193 0.215874920
93 -1.0688745 0.19687437 0.064852135
```



- 더 이상 제거해야 할 관찰값은 없는 것으로 보임

- 다중공선성 확인

  - VIF 값 출력

```
> library(car)
> vif(fit_m2_1)
       x2        x5
1.117063 1.117063
```

VIF(Variance Inflation Factor)

- 설명변수 $X_1, \cdots, X_p$ 사이의 선형 연관성 정도를 파악하기 위한 통계량
- $VIF_i = \frac{1}{1-R_i^2}$. 단, R_i^2은 회귀모형 $X_i = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p + \varepsilon$의 결정계수
- $VIF_1 = 10$이라면 변수 X_1의 변량이 나머지 변수의 선형결합으로 90% 설명이 된다는 의미. 다중공선성이 매우 심각할 수 있음.

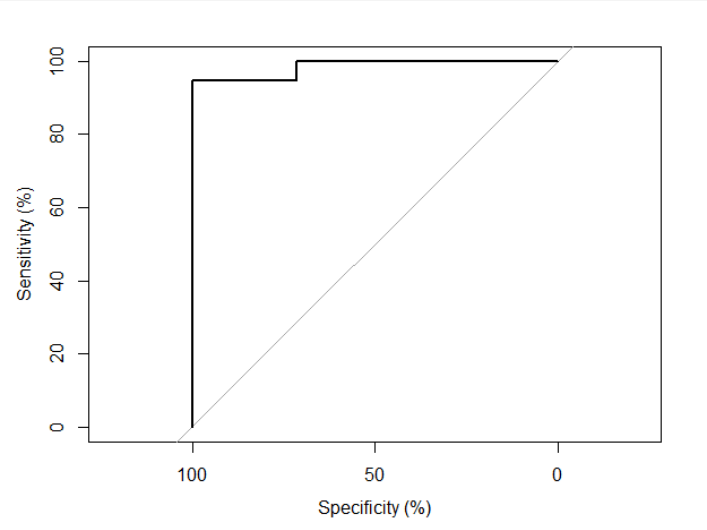
- 예측

```
> fit_m2_1

Coefficients:
(Intercept)           x2           x5
     19.436      -13.699       -3.479

Degrees of Freedom: 128 Total (i.e. Null);  126 Residual
Null Deviance:     165.6
Residual Deviance: 50.89      AIC: 56.89
```

```
> pred <- predict(fit_m2_1, newdata=p21_t, type="response")
```

- AUC 및 정분류율

```
> pROC::roc(p21_t$y, pred, percent=TRUE, plot=TRUE)

Data: pred in 7 controls (p21_t$y 0) < 19 cases (p21_t$y 1).
Area under the curve: 98.5%
```




```
> CCR(d=0.5, y=p21_t$y, pred=pred)
[1]  0.50000 88.46154
```