4. 로지스틱 회귀모형

3) 모형의 적합도 측정

3. 모형의 적합도 측정

- Deviance: 현재 모형의 적합도를 측정하는 유용한 도구

- 이항반응변수의 경우, deviance로 적합도 측정이 어렵기 때문에 다른 측정 도구의 활용이 요구됨

- 이항반응변수 상황에서 사용 가능한 적합도 측정 도구

1. (유사) R^2
2. 정분류율
3. AIC & BIC
4. ROC curve

1) R^2 (유사 R^2)

- 정규분포 회귀모형에서 모형의 적합도 측정하는 대표적인 도구
- 로지스틱 회귀모형에 적용 가능하게 수정

$$R^2 = \frac{1 - \left(\hat{L}_0 / \hat{L}_c\right)^{2/n}}{1 - \left(\hat{L}_0\right)^{2/n}}$$

  - n: 관찰값 개수
  - $\hat{L}_0$: 절편만 있는 모형의 maximized likelihood
  - $\hat{L}_c$: 현재 모형의 maximized likelihood

- 분자: 두 likelihood 비율의 제곱에 기하평균 효과(1/n 승) 추가
- 분모: $0 \leq R^2 \leq 1$을 만족시키기 위한 보정값
- 모형간의 적합도 비교 용도로 사용 가능

- R에서 R^2 계산: 패키지 pscl의 함수 pR2() 이용
  - 예제: 부인 직업 참여 자료

```
> library(carData)
> fit <- glm(lfp~. , family=binomial, Mroz)
>
> library(pscl)
> round(pR2(fit),3)
      llh  llhNull        G2 McFadden      r2ML      r2CU
-452.633 -514.873  124.480    0.121    0.152    0.204
```

  - llh : $\log \hat{L}_c$
  - llhNull : $\log \hat{L}_0$
  - G2 : $-2\left[\log \hat{L}_0 - \log \hat{L}_c\right]$
  - McFadden, r2ML, r2CU : 다른 유형의 유사 R^2

  - 우리가 사용할 통계량: r2CU=0.204

2) 정분류율(Correct Classification Rate: CCR)

- 추정된 $\hat{\pi}(\boldsymbol{x})$로 각 case를 다음의 방법에 의하여 두 범주로 분류

$$\hat{Y} = \begin{cases} 0 & \text{if} \quad \hat{\pi}(\boldsymbol{x}) < d \\ 1 & \text{if} \quad \hat{\pi}(\boldsymbol{x}) \geq d \end{cases}$$

- 관측값 Y와 예측값 $\hat{Y}$으로 2차원 분할표 작성

		Prediction		Total
		$\hat{Y}=1$	$\hat{Y}=0$	
Observation	$Y=1$	n_{11}	n_{10}	n_{1+}
	$Y=0$	n_{01}	n_{00}	n_{0+}
Total		n_{+1}	n_{+0}	n

- CCR 정의

$$CCR = \frac{n_{11} + n_{00}}{n} \times 100$$

  - 현재 모형에 의한 예측값이 관찰값과 동일한 범주로 분류된 비율
  - 모형의 적합 정도를 판단할 수 있는 중요한 도구

- CCR의 문제점

  - 관측비율이 높은 범주로 단순 분류하더라도 정분류율은 50% 이상이 됨.

  - 예: $n_{1+} = 100$, $n_{0+} = 50$인 경우, $n_{+1} = 150$, $n_{+0} = 0$으로 단순 분류
    $CCR = (100/150) \times 100 = 67\%$

- 수정된 CCR

  - 단순 추측으로 보장된 정분류율($\max_r n_{r+}/n$)을 차감

  - $\max_r n_{r+}$: 행 합계 중 최대값. 관측비율이 더 높은 범주의 관측값 개수

  - 수정된 CCR의 정의

$$CCR_{adj} = \left(\frac{n_{11} + n_{00} - \max_r n_{r+}}{n} \right) \times 100$$

  - 모형에 의하여 분류했을 경우, 단순 추측에 의한 분류보다 얼마나 오류를 감소시킬 수 있는지 측정

- 예제 3.4: 부인 직업 참여 자료

  분류 기준값을 d=0.5로 하여 분류표 작성하고 정분류율 계산

```
> library(carData)
> fit <- glm(lfp~. , family=binomial, Mroz)
```

```
> library(dplyr)
> table1 <- mutate(Mroz,
                   lfp_hat=if_else(fit$fitted>=0.5, "yes", "no")) %>%
              with(., table(lfp, lfp_hat))
> addmargins(table1)
     lfp_hat
lfp    no yes Sum
  no  180 145 325
  yes  86 342 428
  Sum 266 487 753
```

- CCR 계산

```
> sum(diag(table1))/sum(table1)*100
[1] 69.32271
```

- 수정된 CCR 계산

```
> y_max <- max(addmargins(table1,2))

> (sum(diag(table1))-y_max)/sum(table1)*100
[1] 12.4834
```

모형 추정에 사용된 자료를 대상으로 계산된 정분류율로
모형의 예측력을 측정하는 것에는 한계가 있음

3) AIC, BIC

- 모형의 적합도 비교에 사용되는 척도

  - Akaike's Information Criterion: $AIC = -2\log\hat{L}_c + 2(p+1)$

  - Bayesian Information Criterion: $BIC = -2\log\hat{L}_c + \log n\,(p+1)$

    $\hat{L}_c$: 현재 모형의 maximized likelihood
    p: 현재 모형에 포함된 설명변수 개수
    n: 자료 개수

- 설명변수의 개수 증가 → $-2\log\hat{L}_c$ 감소
  → $2(p+1)$, $\log(n)(p+1)$ 증가

- AIC, BIC 값이 작을수록 적합도가 더 높다고 판단

예제 3.5: 부인 직업 참여 자료

- 모형 M_1 : 설명변수 $(X_1, X_2, \cdots, X_7)$ 모두 포함
  모형 M_2 : 설명변수 (X_1, X_3, X_4)만 포함

```
> fit <- glm(lfp ~ . , family=binomial, Mroz)
> fit_m2 <- glm(lfp ~ k5 + age + wc,family=binomial, Mroz)
```

- 두 모형의 AIC 비교

```
> fit$aic
[1] 921.2659
> fit_m2$aic
[1] 950.8633

> AIC(fit); AIC(fit_m2)
[1] 921.2659
[1] 950.8633

> BIC(fit); BIC(fit_m2)
[1] 958.2584
[1] 969.3596
```

더 작은 값의 AIC, BIC를 갖는 모형 M_1이 더 선호됨

4) ROC curve

- 추정된 $\hat{\pi}(\boldsymbol{x})$로 각 case를 다음의 방법에 의하여 두 범주로 분류

$$\hat{Y} = \begin{cases} 0 & \text{if} \quad \hat{\pi}(\boldsymbol{x}) < d \\ 1 & \text{if} \quad \hat{\pi}(\boldsymbol{x}) \geq d \end{cases}$$

- ROC(Receiver Operating Characteristic) curve: 주어진 모형의 분류 정확도에 대한 평가 도구

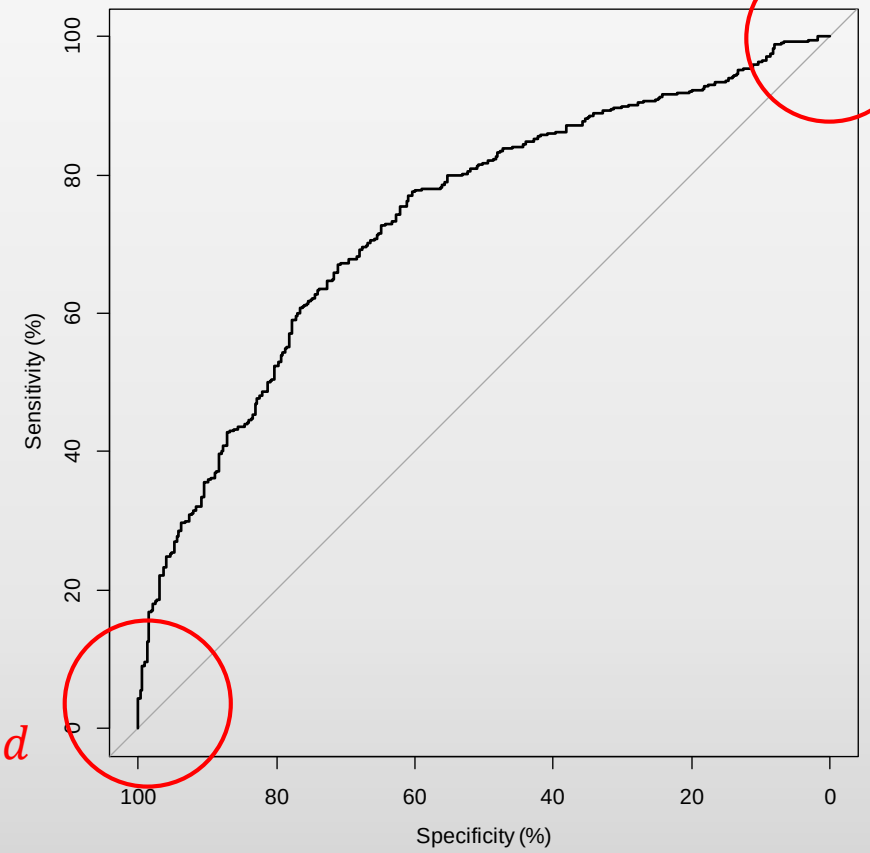
- 분류결과는 분류 기준값 d에 따라 달라짐
- 특정 d에 대한 분류결과가 다음과 같다고 하자.

		Prediction		Total
		$\hat{Y}=1$	$\hat{Y}=0$	
Observation	$Y=1$	n_{11}	n_{10}	n_{1+}
	$Y=0$	n_{01}	n_{00}	n_{0+}
Total		n_{+1}	n_{+0}	n

- Sensitivity(민감도): n_{11}/n_{1+}
  $Y=1$로 관측된 자료 중 $\hat{Y}=1$로 분류된 자료의 비율

- Specificity(특이도): n_{00}/n_{0+}
  $Y=0$로 관측된 자료 중 $\hat{Y}=0$로 분류된 자료의 비율

- ROC curve: 모든 분류 기준값 d에 대하여 specificity(n_{00}/n_{0+})와 sensitivity(n_{11}/n_{1+})를 (x, y) 좌표로 하여 작성된 도표



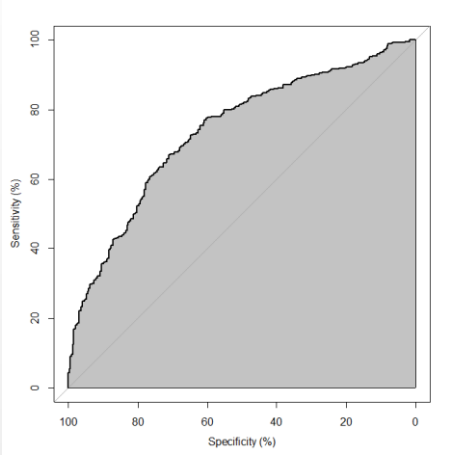
small d

small d
- 대부분의 경우 $\hat{Y} = 1$으로 분류
- n_{11} 증가

large d
- 대부분의 경우 $\hat{Y} = 0$으로 분류
- n_{00} 증가

large d

- AUC(Area under the curve): ROC 곡선 아래 부분의 면적



  - 분류 정확도를 나타내는 척도

  - 급격한 기울기의 ROC curve: 분류 정확도가 높은 모형
    → 큰 값의 AUC

  - Correct 분류 비율 = Wrong 분류 비율 모형:
    → specificity(%)+sensitivity(%)=100% 유지
    → AUC: 50%
    → 분별력이 없는 모형

예제: 부인 직업 참여

- 설명변수를 모두 포함시킨 모형의 ROC curve를 작성하고 AUC 값 계산
- R에서 ROC curve 작성: 패키지 pROC의 함수 roc( ) 이용

```
roc( response, predictor, percent=TRUE, plot=TRUE,
        ci=TRUE)
```

- response: 이항 반응변수 벡터
- predictor: $\hat{\pi}(\boldsymbol{x})$ 벡터. 함수 predict( )로 계산
- percent=TRUE: sensitivities, specificities, AUC 값을 백분율로 계산
- plot=TRUE: ROC curve 작성

```
> library(pROC)
> library(carData)

> fit <- glm(lfp ~ .,family=binomial, Mroz)

> pred <- predict(fit, type="response")

> roc(Mroz$lfp, pred, percent=TRUE, plot=TRUE)


Data: pred in 325 controls (Mroz$lfp no) < 428 cases (Mroz$lfp
yes).
Area under the curve: 73.64%
```

표 3.11과 비교

- ROC curve

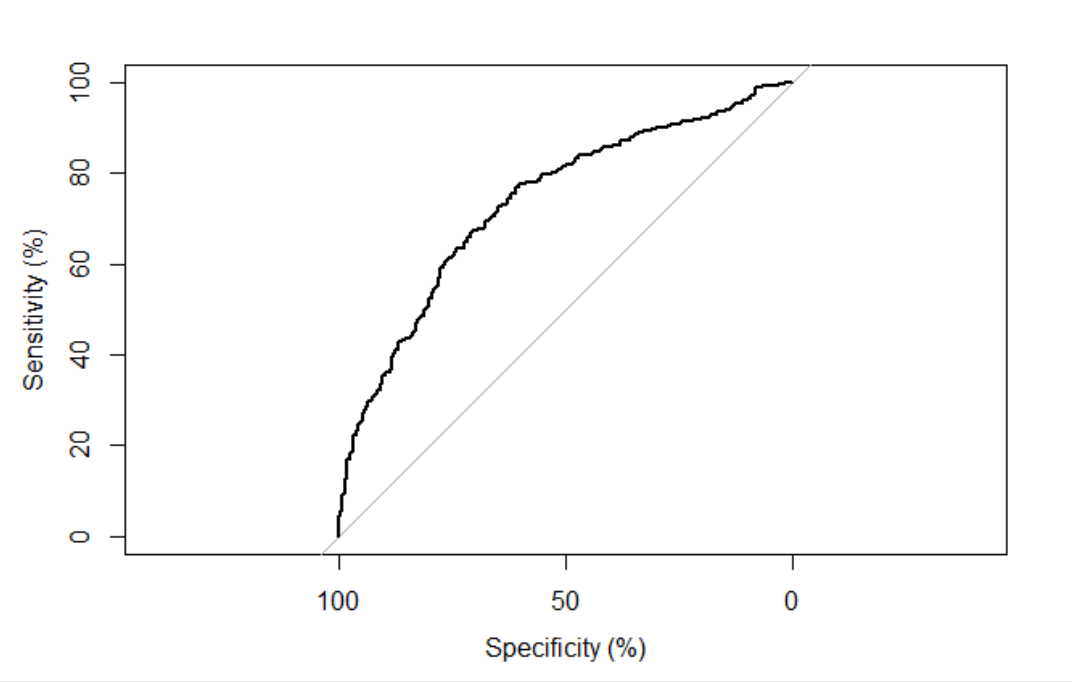


그림 3.1과 비교

연습문제 3.1

- 설명변수: 130개 기업 채권; 투자성 등급(Y=1), 투기성 등급(Y=0)
- 설명변수

  X_1: 총자산 규모(1억 달러)

  X_2 : 레버리지 척도(장기부채/총자본)

  X_3 : 수익성 척도(순이익/총자산)

  X_4 : 불안정척도(순이익 변동계수)

  X_5 : 주식 등급(1~6 척도)
- 자료 파일: p2-1.dat

1) 개별 회귀계수에 대한 유의성을 Wald test로 실시하라.

```
> fit <- glm(y~.-id, family=binomial, data=p21)
Warning message:
glm.fit: 적합된 확률값들이 0 또는 1 입니다

> summary(fit)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.464e+01  3.462e+00   4.228 2.36e-05 ***
x1           3.239e-04  1.691e-04   1.915 0.055471 .
x2          -1.100e+01  3.790e+00  -2.901 0.003715 **
x3           9.126e-01  8.977e-01   1.017 0.309351
x4          -1.387e-01  9.284e-02  -1.494 0.135098
x5          -2.679e+00  7.238e-01  -3.701 0.000215 ***
```

2) $H_0: \beta_2 = \beta_5$
생략

3) $H_0: \beta_3 = \beta_4 = 0$

```
> fit_r <- update(fit, .~.-x3-x4)

> anova(fit_r, fit, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ x1 + x2 + x5
Model 2: y ~ (id + x1 + x2 + x3 + x4 + x5) - id
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       126     59.897
2       124     56.215  2   3.6824   0.1586
```

귀무가설을 기각할 수 없음

4) $H_0: \beta_1 = \beta_2 = \cdots = \beta_5 = 0$

```
> fit_n <- glm(y~1, family=binomial, data=p21)

> anova(fit_n, fit, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ (id + x1 + x2 + x3 + x4 + x5) - id
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       129    167.709
2       124     56.215  5   111.49 < 2.2e-16 ***
```

귀무가설 기각

5) 분류기준값 $d = 0.5$로 분류하고 정분류율 계산

분류행렬

```
> y_hat <- (fit$fitted >= 0.5)*1
> my_table <- table(p21$y, y_hat)

> addmargins(my_table)
     y_hat
        0   1 Sum
  0    39   6  45
  1     5  80  85
  Sum  44  86 130
```

정분류율

수정된 정분류율

```
> sum(diag(my_table))/sum(my_table)*100
[1] 91.53846

> y_max <- max(table(p21$y))
> (sum(diag(my_table))-y_max)/sum(my_table)*100
[1] 26.15385
```

6) 모형 M_1: 설명변수 $(X_1, \cdots, X_5)$. 모형 M_2 : 설명변수 (X_2, X_5)
두 모형의 AIC 비교

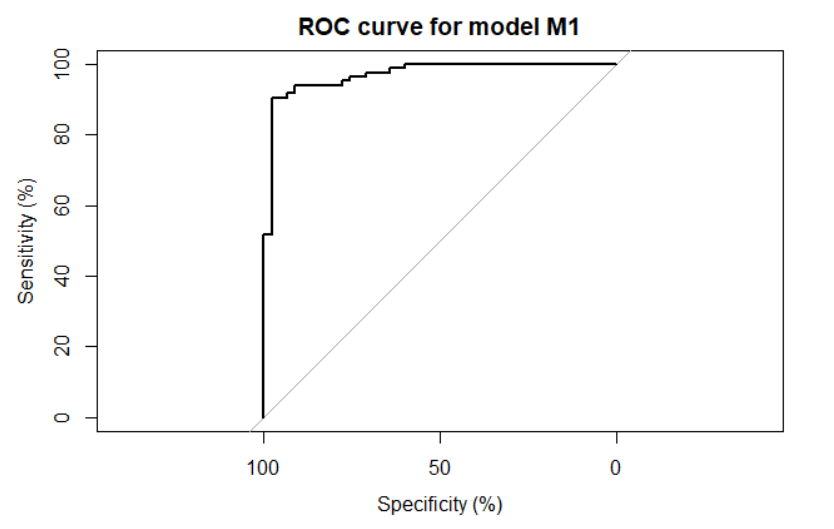
```
> fit_25 <- glm(y~x2+x5,family=binomial,data=p21)

> fit$aic
[1] 68.21485

> fit_25$aic
[1] 68.67355
```

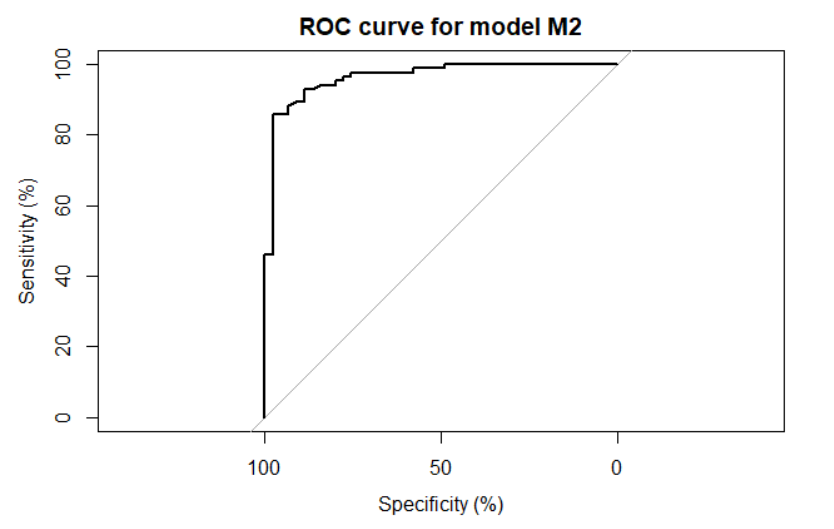
7) 모형 M_1과 M_2에 대한 ROC curve 작성과 비교

```
> library(pROC)
> pred_m1 <- predict(fit, type="response")
> pred_m2 <- predict(fit_25, type="response")
```

```
> roc(p21$y, pred_m1, percent=TRUE, plot=TRUE,
      main="ROC curve for model M1")

Data: pred_m1 in 45 controls (p21$y 0) < 85 cases (p21$y 1).
Area under the curve: 97.07%
```

```
> roc(p21$y, pred_m2, percent=TRUE, plot=TRUE,
      main="ROC curve for model M2")

Data: pred_m2 in 45 controls (p21$y 0) < 85 cases (p21$y 1).
Area under the curve: 96.42%
```





- 두 모형 사이에 큰 차이가 없음
- 단순한 M_2 모형이 더 선호됨

연습문제 3.4

- 기업 파산에 대한 통계적 예측모형 개발
- 반응변수: Y=1 (정상 기업), Y=0 (파산 기업)
- 설명변수
  - X_1: 현금흐름 대 총부채 비율(유동성 지표)
  - X_2: 순이익 대 총자산 비율(수익성 지표)
  - X_3: 유동자산 대 유동부채 비율(단기지급능력 지표)
  - X_4: 유동자산 대 순매출액 비율(자산운영 효율성 지표)
- 자료 파일: p2-4.dat (파산 선고된 21개 기업과 정상 운영되는 25개 기업 자료)

1) 개별 회귀계수에 대한 유의성 검정

```
> fit <- glm(y ~ . -id, family=binomial, data=p24)
> summary(fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.311      2.366  -2.245  0.02477 *
x1             7.060      5.981   1.180  0.23787
x2            -3.509     13.622  -0.258  0.79672
x3             3.414      1.204   2.835  0.00458 **
x4            -2.982      3.068  -0.972  0.33101
```

2) H_0: $\beta_1 = \beta_4$ 생략

3) H_0: $\beta_1 = \beta_2 = \beta_4 = 0$

```
> fit_r <- glm(y ~ x3, family=binomial, data=p24)

> anova(fit_r, fit, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ x3
Model 2: y ~ (id + x1 + x2 + x3 + x4) - id
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        44     35.344
2        41     27.467  3   7.8768  0.04863 *
```

귀무가설 기각

4) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

```
> fit_n <- glm(y ~ 1,family=binomial,data=p24)

> anova(fit_n, fit, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ (id + x1 + x2 + x3 + x4) - id
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        45     63.421
2        41     27.467  4   35.954 2.957e-07 ***
```

- 귀무가설 기각
- 3)의 가설이 기각되었기 때문에 큰 의미 없음
- 가장 먼저 검정해야 할 가설

5) 분류 기준값이 0.5일 때 분류행렬 작성하고 CCR 및 수정된 CCR 계산

```
> y_hat <- (fit$fitted >= 0.5)*1
> my_table <- table(p24$y,y_hat)

> addmargins(my_table)
     y_hat
       0  1 Sum
  0   18  3  21
  1    1 24  25
  Sum 19 27  46
```

```
> sum(diag(my_table))/sum(my_table)*100
[1] 91.30435
```

```
> y_max <- max(table(p24$y))
> (sum(diag(my_table))-y_max)/sum(my_table)*100
[1] 36.95652
```

6) 모형 $M_1(X_1, X_2, X_3, X_4)$ vs 모형 $M_2(X_1, X_3)$ AIC 비교

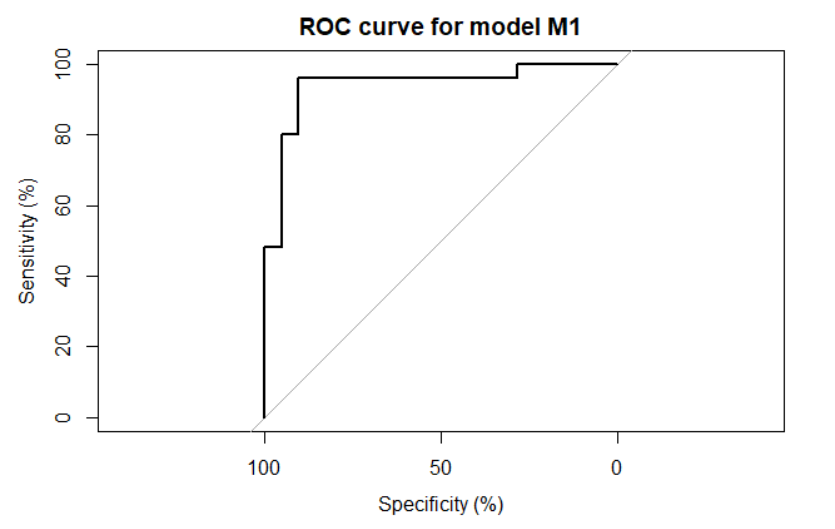
```
> fit_13 <- glm(y~x1+x3, family=binomial, data=p24)

> fit$aic
[1] 37.46722

> fit_13$aic
[1] 34.6579
```

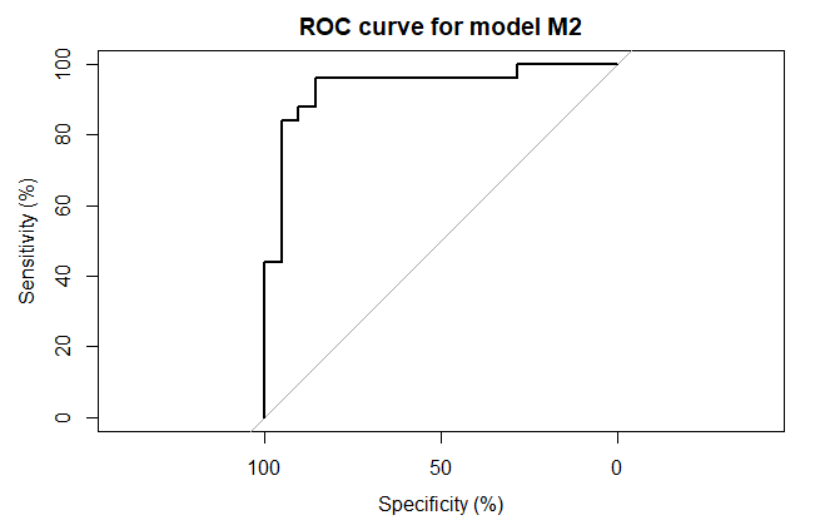
7) 모형 M_1과 M_2의 ROC 곡선 비교

```
> library(pROC)
> pred_m1 <- predict(fit, type="response")
> pred_m2 <- predict(fit_13,type="response")
```

```
> roc(p24$y,pred_m1,percent=TRUE,plot=TRUE,
      main="ROC curve for model M1")

Data: pred_m1 in 21 controls (p24$y 0) < 25 cases (p24$y 1).
Area under the curve: 94.1%
```

```
> roc(p24$y,pred_m2,percent=TRUE,plot=TRUE,
      main="ROC curve for model M2")

Data: pred_m2 in 21 controls (p24$y 0) < 25 cases (p24$y 1).
Area under the curve: 93.71%
```


ROC curve for model M1
Sensitivity (%)
0
20
40
60
80
100
100
50
0
Specificity (%)


ROC curve for model M2
Sensitivity (%)
0
20
40
60
80
100
100
50
0
Specificity (%)

연습문제 3.5

- 신생아 몸무게(BWT)가 2.5kg 미만인 미숙아 출산과 관련된 위험요인에 관한 연구
- 반응변수: Low=0(정상), Low=1(미숙아)
- 설명변수

  Age: 출산 시 산모 나이

  Lwt: 출산 직전 산모 몸무게

  Race: 인종(1=백인, 2=흑인, 3=기타)

  Smoke: 임신 중 흡연여부(1=흡연, 0=금연)

  Ptl: 미숙아 출산 경험여부(1=있음, 0=없음)

  Ht: 고혈압여부(1=고혈압, 0=정상)

  Ui: 자궁자극성여부(1=있음, 0=없음)

  Ftv: 임신 첫 3개월간 진찰횟수

자료 입력

```
> p25 <- read.table("D:/Data/p2-5.dat")

> names(p25) <- c("id","low","age","lwt","race","smoke",
                  "ptl","ht","ui","ftv","bwt")

> p25 <- dplyr::mutate(p25, race=factor(race))
```

- 변수 age와 lwt를 제외한 나머지 변수는 모두 범주형
- 변수 race는 3개 범주: factor로의 전환이 필수
- 나머지 변수는 모두 2개 범주: 0 또는 1의 값을 갖고 있기 때문에 factor로 전환하지 않아도 함수 glm( )에서 사용 가능

1) 개별 회귀계수에 대한 유의성 검정

```
> fit <- glm(low~.-id-bwt, family=binomial, p25)
> summary(fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.480623   1.196888   0.402  0.68801
age         -0.029549   0.037031  -0.798  0.42489
lwt         -0.015424   0.006919  -2.229  0.02580 *
race2        1.272260   0.527357   2.413  0.01584 *
race3        0.880496   0.440778   1.998  0.04576 *
smoke        0.938846   0.402147   2.335  0.01957 *
ptl          0.543337   0.345403   1.573  0.11571
ht           1.863303   0.697533   2.671  0.00756 **
ui           0.767648   0.459318   1.671  0.09467 .
ftv          0.065302   0.172394   0.379  0.70484
```

race2 (1=흑인, 0=그 외), race3(1=기타, 0=그 외) → reference 범주가 백인

- 다중 범주를 갖고 있는 factor가 설명변수가 된 경우의 해석 방법

  - 변수 race는 3개의 범주를 갖고 있는 factor
  - 이 경우 race를 설명변수에 포함시키면 2개의 dummy 변수가 생성됨

  - race2는 흑인, race3는 기타 인종의 경우에만 각각 1이 됨
  - race2는 흑인과 백인의 차이, race3는 기타 인종과 백인의 차이

  - 이 경우 race의 첫 번째 범주인 "1"(백인)을 reference 범주라고 함

    교재에서는 reference 범주를 "3"(기타 인종)으로 처리하여
    D_1(1=백인, 0=그 외), D_2(1=흑인, 0=그 외)를 dummy 변수로 사용함

- reference 범주를 변경하고 다시 적합

```
> p25 <- dplyr::mutate(p25,race=relevel(race, ref="3"))

> fit <- glm(low ~ . -id -bwt, family=binomial, p25)

> summary(fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.361119   1.104589   1.232  0.21786
age         -0.029549   0.037031  -0.798  0.42489
lwt         -0.015424   0.006919  -2.229  0.02580 *
race1       -0.880496   0.440778  -1.998  0.04576 *
race2        0.391764   0.537609   0.729  0.46618
smoke        0.938846   0.402147   2.335  0.01957 *
ptl          0.543337   0.345403   1.573  0.11571
ht           1.863303   0.697533   2.671  0.00756 **
ui           0.767648   0.459318   1.671  0.09467 .
ftv          0.065302   0.172394   0.379  0.70484
```

race1 백인과 기타 인종의 차이, race2 흑인과 기타 인종의 차이

4) H_0: $\beta_1 = \cdots = \beta_9 = 0$

```
> fit_n <- glm(low ~ 1,family=binomial, p25)

> anova(fit_n, fit, test="Chisq")
Analysis of Deviance Table

Model 1: low ~ 1
Model 2: low ~ (id + age + lwt + race + smoke + ptl + ht + ui + ftv +
    bwt) - id - bwt
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       188     234.67
2       179     201.28  9   33.387 0.0001143 ***
```

- 귀무가설 기각
- 적어도 한 개 이상의 유의한 변수 존재

3) H_0: $\beta_{\text{Age}} = \beta_{\text{Ptl}} = \beta_{\text{Ftv}} = 0$

```
> fit_r <- update(fit, . ~ . - age - ptl - ftv)
> anova(fit_r, fit, test="Chisq")
Analysis of Deviance Table

Model 1: low ~ lwt + race + smoke + ht + ui
Model 2: low ~ (id + age + lwt + race + smoke + ptl + ht + ui + ftv +
    bwt) - id - bwt
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       182     204.22
2       179     201.28  3   2.9318   0.4023
```

귀무가설 기각할 수 없음

5-1) 분류 기준값이 0.05일 때 분류행렬 작성하고 CCR과 수정된 CCR 계산

```
> low_hat <- (fit$fitted >= 0.05)*1

> my_table <- table(p25$low, low_hat)

> addmargins(my_table)
     low_hat
        0   1 Sum
  0     5 125 130
  1     0  59  59
  Sum   5 184 189

> sum(diag(my_table))/sum(my_table)*100
[1] 33.86243

> y_max <- max(table(p25$low))
> (sum(diag(my_table))-y_max)/sum(my_table)*100
[1] -34.92063
```

5-2) 분류 기준값이 0.3, 0.4, 0.6, 0.7인 경우 같은 분석 실시
어떤 기준값에서 정분류율이 가장 높은가?

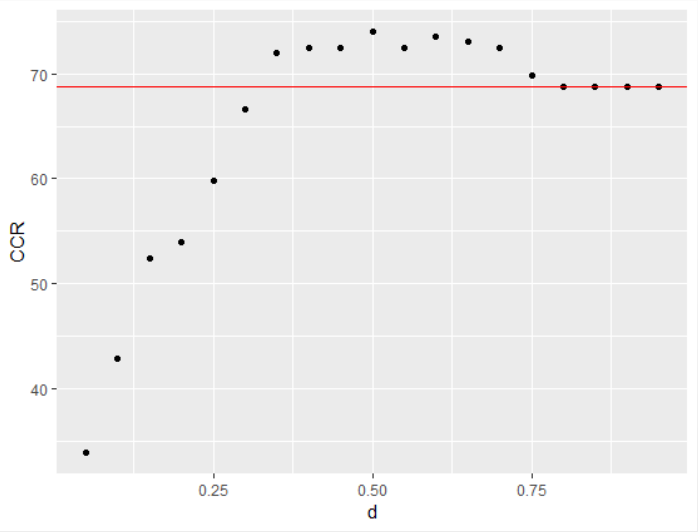
방법 1: 앞에서 실행시킨 code에서 d값만 수정하고 반복 시행
방법 2: 사용자 정의함수를 만들어서 시행
→ 동일한 분석을 반복 시행해야 하는 경우 효과적인 방법

- 함수 CCR 작성 및 실행

```
> 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=0.05, y=p25$low, pred=fit$fitted)
[1]  0.05000 33.86243
```

- 다양한 d 값에 대한 CCR 계산 및 그래프 작성



d=0.5일 때 정분류율이 가장 높게 나옴

- 다양한 d 값에 대한 CCR 계산

```
> d <- seq(0.05,0.95,by=0.05)
> res <- numeric(length(d)*2)
> dim(res) <- c(length(d),2)

> for(i in seq_along(d)){
    res[i,]=CCR(d=d[i], y=p25$low, pred=fit$fitted)
  }
```

- 계산 결과 그래프 작성

```
> colnames(res) <- c("d","ccr")

> max_t <- max(table(p25$low))/sum(table(p25$low))*100

> library(ggplot2)

> ggplot(data.frame(res), aes(x=d, y=ccr)) +
    geom_point() +
    geom_hline(yintercept=max_t, color="red") +
    labs(x="d", y="CCR")
```

6) 모형 M_1: 모든 설명변수 포함
모형 M_2: 설명변수 (Lwt, Race, Smoke, Ht, Ui)만 포함

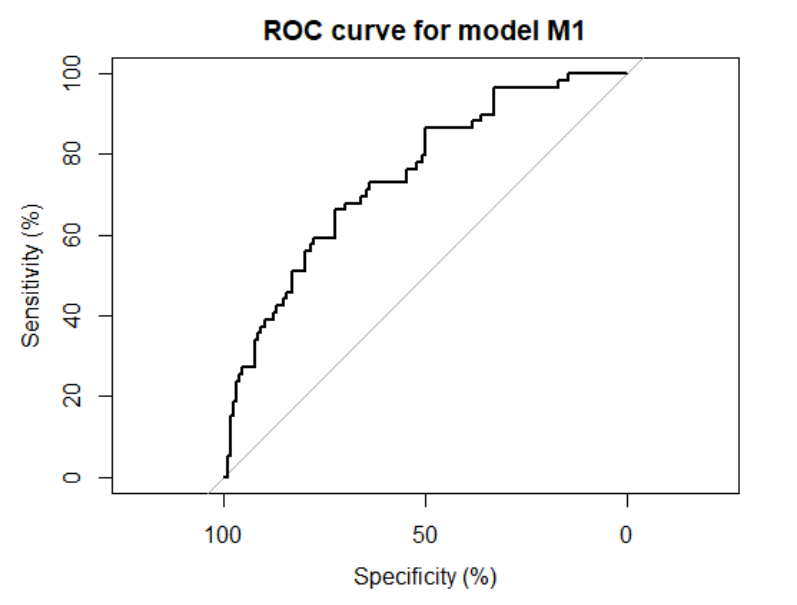
두 모형의 AIC 비교

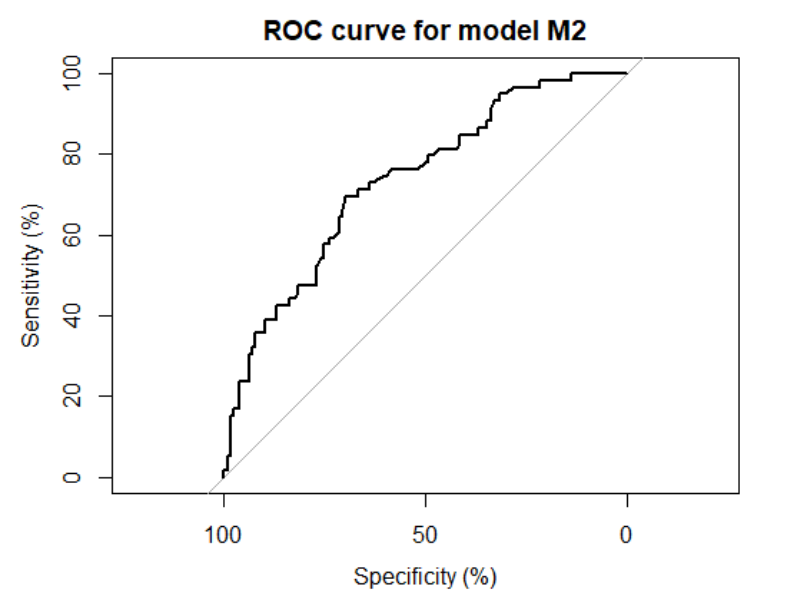
```
> fit_m1 <- glm(low ~ . -id -bwt, family=binomial, p25)
> fit_m2 <- update(fit_m1, . ~ . -age -ptl -ftv)

> fit_m1$aic
[1] 221.2848
> fit_m2$aic
[1] 218.2166
```

7) 모형 M_1과 M_2의 ROC 곡선 작성

```
> library(pROC)

> pred_m1 <- predict(fit_m1, type="response")
> pred_m2 <- predict(fit_m2, type="response")
```

```
> roc(p25$low, pred_m1, percent=TRUE, plot=TRUE,
      main="ROC curve for model M1")

Data: pred_m1 in 130 controls (p25$low 0) < 59 cases (p25$low 1).
Area under the curve: 74.62%

> roc(p25$low, pred_m2, percent=TRUE, plot=TRUE,
      main="ROC curve for model M2")

Data: pred_m2 in 130 controls (p25$low 0) < 59 cases (p25$low 1).
Area under the curve: 73.51%
```


ROC curve for model M1
Sensitivity (%)
0
20
40
60
80
100
100
50
0
Specificity (%)


ROC curve for model M2
Sensitivity (%)
0
20
40
60
80
100
100
50
0
Specificity (%)