# 2장. 회귀분석

최호식

경기대학교 응용통계학과

Mar, 2018(Kyonggi University)

# 학습내용

1. 개요
2. 단순선형회귀
3. 다중선형회귀
4. 다항회귀
5. 범주형 입력변수
6. 로지스틱 회귀

# 회귀분석 I

- 연속형 출력변수 $Y$와 입력변수 $X$의 함수관계를 모형화:

$$Y = f(X) + \epsilon,$$

  $\epsilon$은 오차항으로서 기대값이 0이고, 분산이 $\sigma^2$임

- $X$가 주어졌을 때, $Y$의 조건부 기대값을 특정형태의 함수로 가정하고 자료로부터 그 함수를 추정
  - 선형: $f$가 선형함수
    (예) $f(x) = \sin(2x)$이면 비선형 모형
  - 단순: $X$가 1차원, 다중: $X$가 $p \geq 2$차원

# 단순선형회귀 I

- 모형

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \cdots, n.$$

  $\epsilon_i$: 서로 독립적으로 평균이 0이고 분산이 $\sigma^2$인 오차항

- 추정
  - 오차제곱합 $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$을 최소화
  - 최소제곱 추정치(least square estimate)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- 검정
  - 적합된 추정식이 의미가 있는지를 검정하는 것은 회귀계수 $\beta_1$이 0인지를 검정하는 것과 동일

# 단순선형회귀 II

- 검정통계량은

$$t = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}$$

- $|t|$가 크면 $\beta_1$이 0이라는 가설(귀무가설)을 기각(추정된 회귀식이 유의함).

- 예측 및 해석
  - 새로운 입력변수값 $x$의 값이 주어지면 출력변수 $y$의 예측값은

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

  - $\hat{\beta}_j$는 $x$가 한 단위 증가할 때의 $y$의 증가량을 의미.

- 제곱합의 분해

$$\begin{aligned} \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= = SSR + SSE \end{aligned}$$

# 단순선형회귀 III

- 결정계수
  - $R^2 = SSR/SST, 0 \leq R^2 \leq 1$
  - 1에 가까울수록 회귀모형이 자료를 잘 설명

- 선형 회귀모형에 대한 세가지 가정
  - 선형성: 입력변수와 출력변수의 관계가 선형적
  - 등분산성: 오차의 분산이 입력변수와 무관하게 일정
  - 정규성: 오차의 분포가 정규분포를 따름

# 단순선형회귀 IV

- 가정에 대한 검증
  - 단순선형회귀모형의 경우 입력변수값과 출력변수값의 산점도, 다중선형회귀모형에서는 잔차 $(y_i - \hat{y}_i)$와 출력변수의 산점도를 그려보아 선형성을 만족하는지 체크
  - 모든 가정이 만족되는 경우에 오차(혹은 잔차)는 랜덤하므로 어떤 패턴이 발견되면 선형회귀모형의 가정을 의심

- 가정이 맞지 않을 때
  - 선형성: 다항회귀(polynomial regression) 등의 비선형회귀모형이나 회귀모형에 아무런 가정도 하지 않는 의사결정나무, 신경망 모형과 같은 비모수적 회귀모형 등을 고려
  - 등분산성: 가중회귀방법(weighted regression)을 사용
  - 정규성: 로버스트 회귀(robust regression)를 고려

# 단순회귀 예제 I

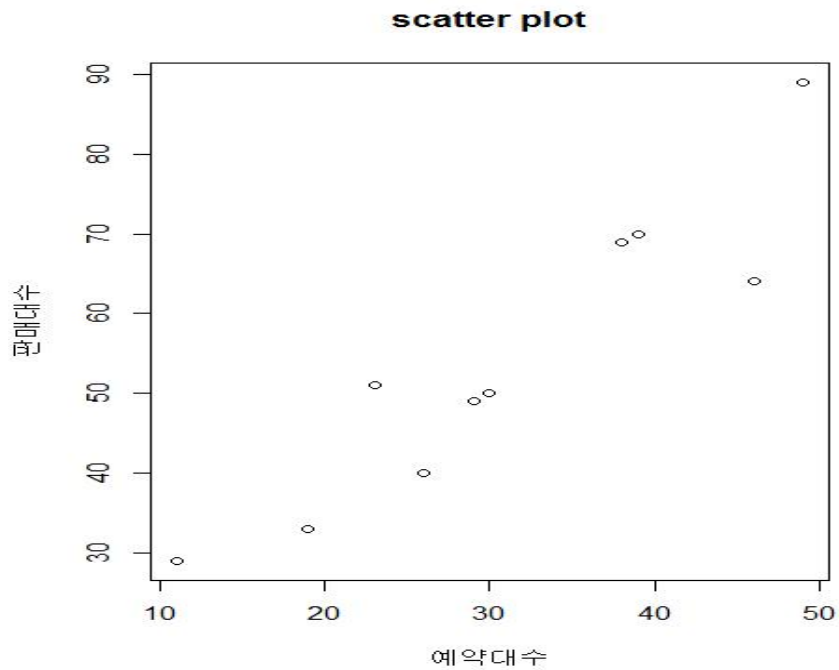
- 에어컨 예약대수를 이용하여 에어컨 판매대수를 예측
  - 10년간의 에어컨 예약대수와 판매대수를 관측(1단위: 1,000대)
  - $x_i$는 에어컨 예약대수, $y_i$는 에어컨 판매대수인 단순선형회귀 고려
  - 분산분석표

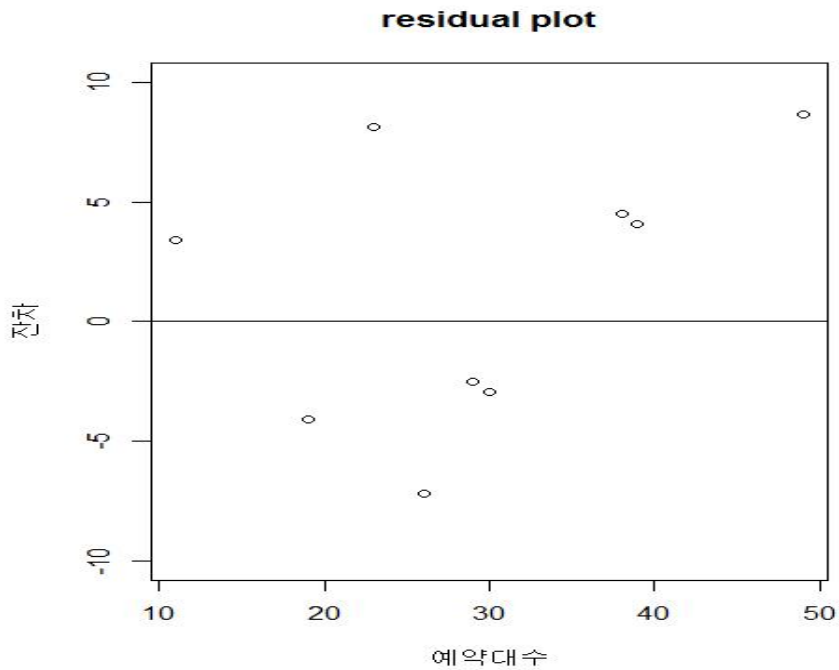
| 변수 | 자유도 | 추정값 | 표준오차 | $t$ | 유의확률 |
|---|---|---|---|---|---|
| 절편 | 1 | 9.7364 | 6.6620 | 1.471 | 0.1796 |
| $x$ | 1 | 1.4407 | 0.2004 | 7.188 | 0.0001 |

- 추정된 모형식이 $y = 9.74 + 1.44x$이고 올해 에어컨 예약대수가 45단위이면 에어컨 판매대수는 $9.74 + 1.44 \times 45 = 74.54$단위로 예측
- 에어컨 예약대수가 1단위 증가하면 에어컨 판매대수는 1.44단위 증가하는 것으로 해석
- $x$에 대한 유의확률(p-value)이 0.0001로 매우 작으므로 회귀계수 $\beta_1$이 0이라는 가설은 기각
- $R^2 = 0.8659$로 자료의 전체 변동중 회귀모형이 86.59%를 설명

# 단순회귀 예제 II

- 산점도와 잔차도




- 산점도: 예약대수와 판매대수 사이의 강한 선형관계가 존재있다.
- 잔차도: 잔차들이 특별한 패턴이 없이 랜덤하게 잘 분포

# 다중선형회귀 I

- 모형

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i.$$

- $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$: 입력변수 벡터, $y_i \in \mathbb{R}$: 출력변수, $\epsilon_i$은 오차항으로 평균이 0이고 분산이 $\sigma^2$
- 훈련자료 $(x_1, y_1), \ldots, (x_n, y_n)$
- 모수 $\beta_0, \beta_1, \ldots, \beta_p$의 추정
- 오차제곱합

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

  를 최소화

# 다중선형회귀 II

- 분산분석표

| 요인 | 제곱합 | 자유도 | 제곱평균 | $F$-값 |
|---|---|---|---|---|
| 회귀 | 회귀제곱합 (SSR) | $p$ | $MSR = SSR/p$ | $F = MSR/MSE$ |
| 오차 | 잔차제곱합 (SSE) | $n-p-1$ | $MSE = SSE/(n-p-1)$ | |
| 계 | 전체제곱합 (SST) | $n-1$ | | |

$F$-값이 크면 귀무가설 $H_0 : \beta_1 = \cdots = \beta_p = 0$ 기각

- $\beta_j$에 대한 유의성 검정 $H_0 : \beta_j = 0$ 대 $H_1 : \beta_j \neq 0$대한 검정통계량은

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

$|t_j|$이 크면 회귀계수 $\beta_j$는 유의하다고 결론을 내림

# 다중선형회귀 III

- 예측 및 해석
  - 새로운 입력변수값 $x_1, \ldots, x_p$에 대한 출력변수 $y$의 예측값은

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots \hat{\beta}_p x_p$$

  - $\hat{\beta}_j$는 다른 변수들의 값이 고정되었을 때 $x_j$가 한 단위 증가할 때의 $y$의 증가량을 의미

# 다중선형회귀 예제 I

- 아파트의 겨울의 난방비 예측
- 20개의 아파트를 임의로 추출하여 겨울철 외부 온도, 단열재의 두께, 창문수, 아파트의 나이를 조사
- 분산분석표

| 요인 | 제곱합 | 자유도 | 제곱평균 | $F$-값 | 유의확률 |
|---|---|---|---|---|---|
| 회귀 | 175042.37 | 4 | 43760.59 | 14.77 | $<$0.001 |
| 오차 | 44443.38 | 15 | 2962.89 | | |
| 계 | 219485.75 | 19 | | | |

- 유의확률이 0.001보다 작으므로 이 회귀모형은 유의
- 회귀계수

| 입력변수 | 회귀계수 | 표준오차 | $t$ | 유의확률 |
|---|---|---|---|---|
| 절편 | 294.46 | 64.61 | 4.55 | $<$ 0.001 |
| 외부.온도 | -8.44 | 1.55 | -5.43 | $<$0.001 |
| 단열재.두께 | -14.90 | 5.30 | -2.80 | 0.013 |
| 창문수 | -1.16 | 5.11 | -0.22 | 0.822 |
| 아파트나이 | 6.26 | 4.31 | 1.45 | 0.166 |

# 다중선형회귀 예제 II

- 외부온도 1도가 올라가면 난방비가 8.44단위 감소
- 단열재의 두께가 1단위 증가하면 난방비가 14.91단위 감소
- 창문수가 1개 증가할수록 난방비가 1.165단위 감소
- 아파트의 나이가 1년 증가하면 난방비가 6.26단위 증가
- 단, 창문수와 아파트의 나이는 p-value가 0.822, 0.166로 통계적으로 유의하지는 않음.

# 최소제곱 추정의 원리* I

- $\mathbf{X}$: $n \times (p+1)$ 설계행렬, $y = (y_1, \ldots, y_n)^T$는 출력변수 벡터
- 오차제곱합

$$\begin{aligned} SSE(\beta) &= (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta) \\ \frac{\partial SSE}{\partial \beta} &= -2\mathbf{X}^T (y - \mathbf{X}\beta) \\ \frac{\partial^2 SSE}{\partial \beta \partial \beta^T} &= 2\mathbf{X}^T \mathbf{X} \end{aligned}$$

- 정규방정식(normal equation): $\frac{\partial SSE}{\partial \beta} = \mathbf{X}^T (y - \mathbf{X}\beta) = 0$
- $\mathrm{rank}\mathbf{X} = p+1$이면 유일한 해 $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$가 존재하며, $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{H}y$임.
- 정사영행렬(projection or hat matrix) $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
  - $\mathbf{H}^2 = \mathbf{H}$를 만족

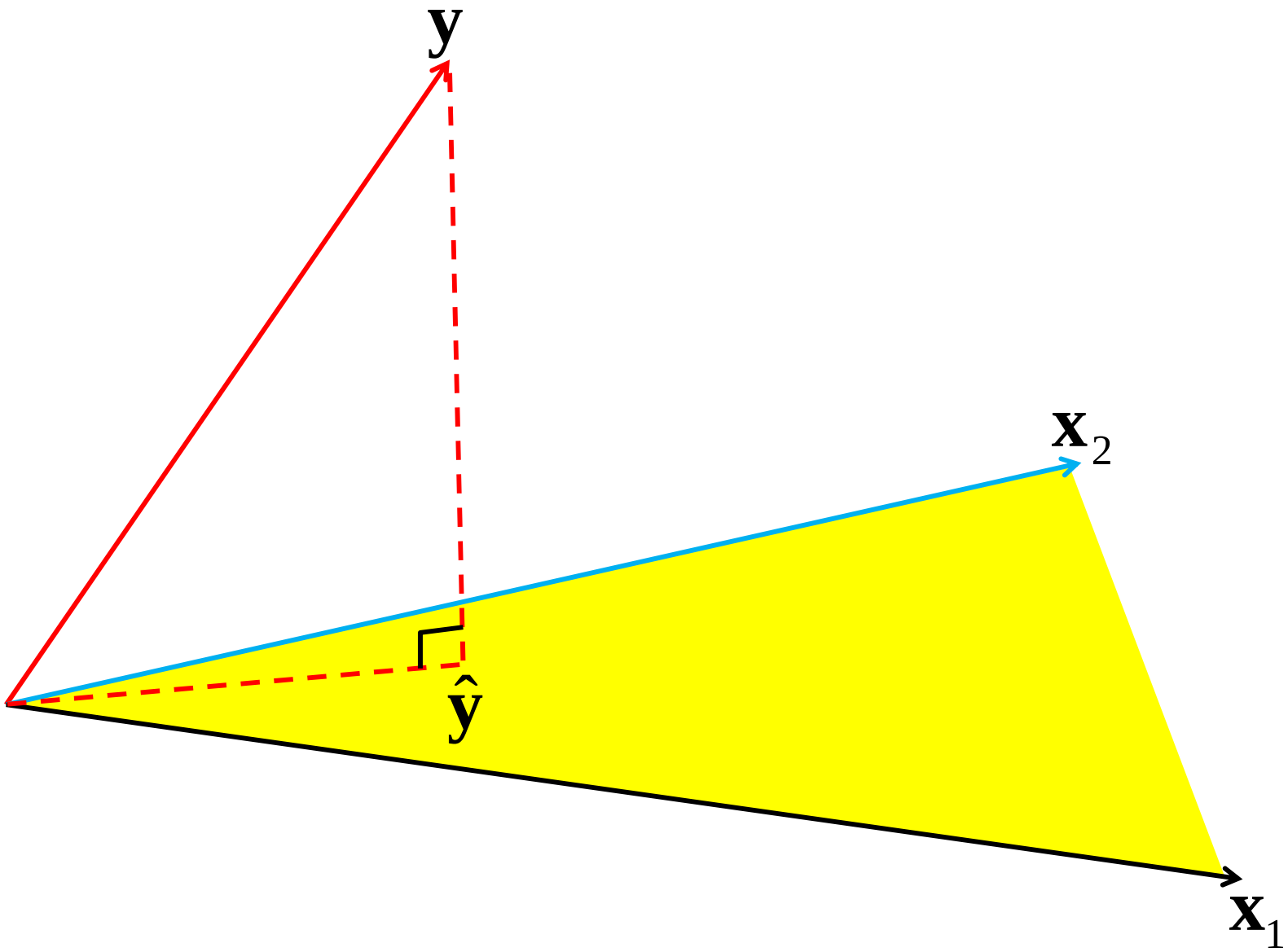
# 최소제곱 추정의 원리* II

- $\mathbf{H}y$는 주어진 벡터 $y$를 $\mathbf{X}$의 열벡터에 의해 생성된 선형공간으로 사영한 값임. 즉, $S = \{\mathbf{X}a : a \in \mathbb{R}^p\}$에 대하여

$$\hat{y} = \arg\min_{z \in S} \sum_{i=1}^{n} (y_i - z_i)^2$$

# 최소제곱 추정의 원리* III

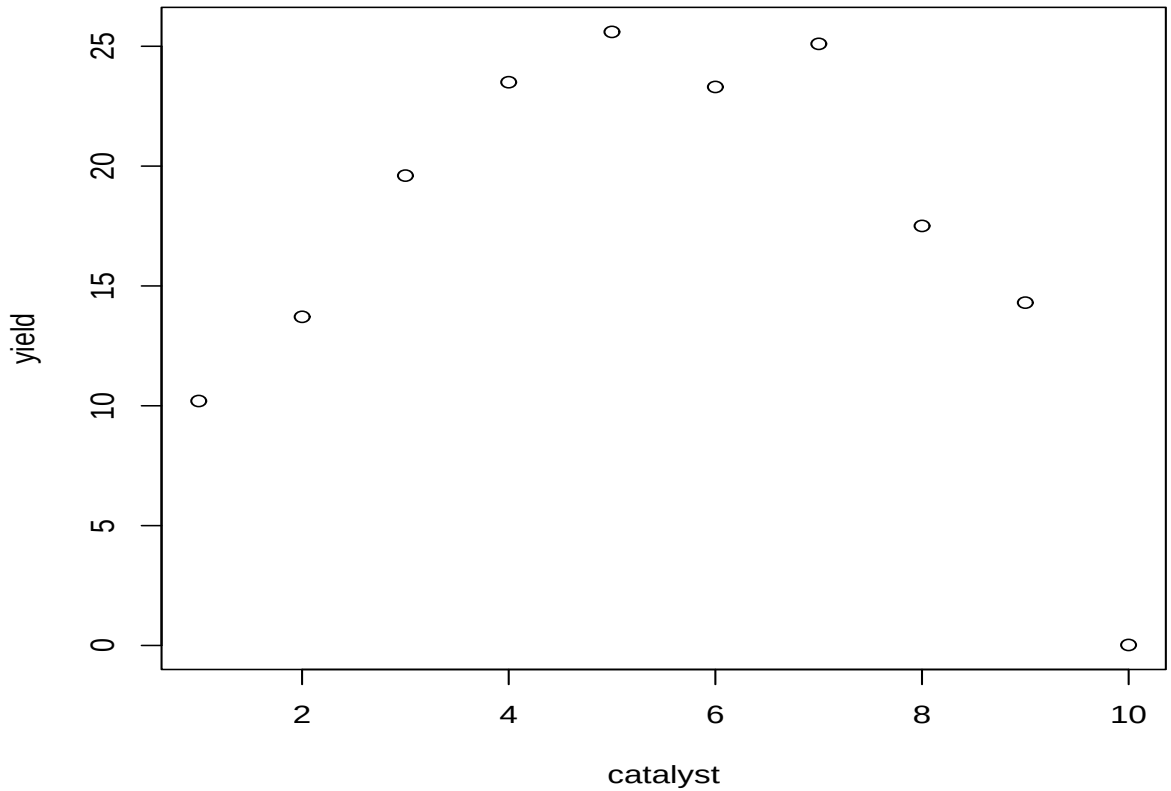
- 최소제곱법의 기하적 해석



# 다항회귀

- 다항회귀모형
$$y = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p + \epsilon$$
- $z_1 = x, \ldots, z_p = x^p$라 놓고 $p$차원의 입력변수 $z_1, \ldots, z_p$와 출력변수 $y$간의 다중선형회귀모형을 적합하여 모수 추정
- 차수 $p$의 결정
  - $p = 2$부터 순차적으로 값을 증가시키면서 모형을 적합. $p$가 증가하면서 결정계수값은 계속해서 증가하므로 결정계수값의 증가가 둔화되는 $p$를 선택함

# 다항회귀 예제 I



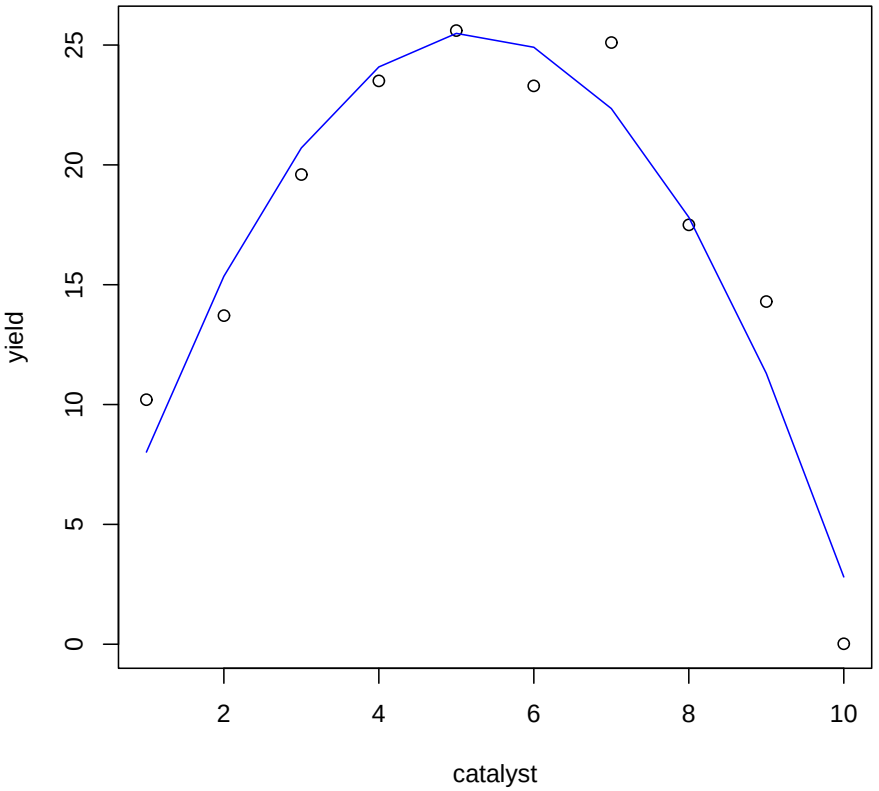

- 산점도

  출력변수와 입력변수의 관계가 선형이 아님

# 다항회귀 예제 II

2차함수를 적합한 결과

| 입력변수 | 회귀계수 | 표준오차 | $t$ | 유의확률 |
|---|---|---|---|---|
| 절편 | -1.26 | 2.71 | -0.46 | 0.65 |
| $x$ | 10.32 | 9.11 | 9.11 | 0.00 |
| $x^2$ | -0.99 | 0.10 | 9.88 | 0.00 |



# 범주형 입력변수 I

- 범주의 수가 $K$개인 범주형 입력변수는 $(K-1)$개의 가변수(dummy variable) $z_1, \ldots, z_{K-1}$로 코딩
- (예) $K = 3$인 경우 다음과 같이 2개의 가변수 $z_1$과 $z_2$를 고려

| 범주 | $z_1$ | $z_2$ |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0 | 1 |
| 3 | 0 | 0 |

- 가변수를 이용한 선형모형은

$$y = \beta_0 + \sum_{k=1}^{K-1} \beta_k z_k + \epsilon$$

# 범주형 입력변수 II

- 회귀계수의 의미
  - $\beta_0$: $K$번째 범주에서의 $y$의 평균
  - $\beta_0 + \beta_k, k \leq K-1$은 $k$번째 범주에서의 $y$의 평균
  - $\beta_k$: $k$번째 범주에서의 $y$의 평균과 $K$번째 범주에서의 $y$의 평균의 차이

계절에 따른 매출액

- 자료

| 봄 | 여름 | 가을 | 겨울 | 봄 | 여름 | 가을 | 겨울 |
|---|---|---|---|---|---|---|---|
| 231.21 | 337.28 | 336.81 | 300.44 | 283.91 | 255.61 | 285.49 | 338.52 |
| 봄 | 여름 | 가을 | 겨울 | 봄 | 여름 | 가을 | 겨울 |
| 238.37 | 259.98 | 280.13 | 306.02 | 224.70 | 264.45 | 317.51 | 327.78 |

- 가변수 $z_1, z_2, z_3$

| 계절변수 | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|
| 봄 | 0 | 0 | 0 |
| 여름 | 1 | 0 | 0 |
| 가을 | 0 | 1 | 0 |
| 겨울 | 0 | 0 | 1 |

# 범주형 입력변수 III

- 추정결과

| 입력변수 | 회귀계수 | 표준오차 | $t$ | 유의확률 |
|---|---|---|---|---|
| 절편 | 244.55 | 14.3 | 17.10 | $<$0.001 |
| z1 | 34.78 | 20.22 | 1.72 | 0.111 |
| z2 | 60.44 | 20.22 | 2.98 | 0.011 |
| z3 | 73.64 | 20.22 | 3.64 | 0.003 |

  - 여름의 매출액은 봄의 매출액 비해 34.78단위 높으나 통계적으로 유의하지 않음
  - 가을의 매출액은 봄의 매출액보다 60.44단위 높으며, 겨울의 매출액은 봄의 매출액에 비해 73.64단위 높음

# 단순 로지스틱 회귀 I

- 출력변수 $y = 0$ 또는 1이고 입력변수가 $x$인 경우에 단순선형회귀모형

$$y = \beta_0 + \beta_1 x + \epsilon$$

을 적용

  - 추정값 $\beta_0 + \beta_1 x$는 범위 $[0, 1]$을 벗어날 수 있음
  - 오차항 $\epsilon$의 분포가 정규분포가 아님

- $x$가 주어졌을때 $Y$의 조건부 평균이라기 보다 조건부 확률을 적절한 연결함수(link function) $p$를 통해 모형화. $p$는 연속인 증가함수로 $[0, 1]$에서 값을 가짐
  - 로지스틱 모형: $p(x) = \exp(x)/(1 + \exp(x))$
  - 검벨(Gumbel) 모형: $p(x) = \exp(-\exp(x))$
  - 프로빗(probit) 모형: $p(x) = \Phi(x)$(표준정규분포의 분포함수), 생물학 독성실험 등에서 사용

# 단순 로지스틱 회귀 II

- 단순 로지스틱 모형:

$$\Pr(Y = 1|x) = p(\beta_0 + \beta_1 x) = \exp(\beta_0 + \beta_1 x)/(1 + \exp(\beta_0 + \beta_0 x))$$

- 오즈비(odds ratio):

$$\begin{aligned} & \frac{\Pr(Y = 1|x+1)/\Pr(Y = 1|x)}{\Pr(Y = 0|x+1)/\Pr(Y = 0|x)} \\ = & \frac{\Pr(Y = 1|x+1)\Pr(Y = 0|x)}{\Pr(Y = 0|x+1)\Pr(Y = 1|x)} \\ = & \exp(\beta_1). \end{aligned}$$

  - 입력변수 $x$가 한단위 증가할 때 오즈비가 일정하게 증가
  - 오즈(odds)는 $\Pr(Y = 1|x)/\Pr(Y = 0|x)$로 정의되며, 입변수 $x + 1$에서의 오즈와 입력변수 $x$에서의 오즈의 비
  - (예) $x$: 소득, $y$: 어떤 상품에 대한 구입여부(1=구입, 0=미구입)

# 단순 로지스틱 회귀 III

- $\hat{\beta}_1 = 3.72$: 소득이 한 단위 증가하면 물품을 구매하지 않을 확률에 대한 구매할 확률의 오즈비가 $\exp(3.72) = 42$배 증가

- 추정
  - 우도함수(likelihood function)는

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(\beta_0 + \beta_1 x_i)^{y_i} (1 - p(\beta_0 + \beta_1 x_i))^{1-y_i}$$

  - 로그 우도함수

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i \log p(\beta_0 + \beta_1 x_i) + (1 - y_i) \log(1 - p(\beta_0 + \beta_1 x_i)))$$

  - 최대우도추정량(maximum likelihood estimator) $(\hat{\beta}_0, \hat{\beta}_1)$을 이용

# 단순 로지스틱 회귀 IV

- 검정
  - 우도비검정통계량

$$\chi^2 = -2\left(\max_{\beta_0} l(\beta_0, 0) - l(\hat{\beta}_0, \hat{\beta}_1)\right)$$

    을 이용
  - $\chi^2$통계량은 근사적으로 자유도가 1인 카이제곱 분포를 따름. 값이 크면 $\beta_1 = 0$는 기각

# 다중 및 다항 로지스틱 회귀로의 확장

- 다항 로지스틱 회귀모형

$$\Pr(Y = 1|x) = p(\beta_0 + \beta_1 x + \cdots + \beta_p x^p)$$

- 다중 로지스틱 회귀모형은 입력변수 $x_1, \ldots, x_p$에 대하여

$$\Pr(Y = 1|x) = p(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

- 모수의 추정: 최대우도법
- 유의성 검정: 우도비 검정
- 입력변수가 범주형인 경우에도 가변수 이용
- 변수선택: 선형모형의 방법을 적용(선택기준은 오차제곱합 대신에 로그우도함수값을 사용)

# 다중 로지스틱 회귀 예제 I

- 회사채의 신용등급을 여러 가지 재무변수를 이용하여 설명
- 출력변수: 회사채 신용등급(안정=1, 위험=0)
- 입력변수: 자산대비 부채현황 지표($x_1$), 현금회전율($x_2$), 종업원수($x_3$)이다.
- 종업원수는 50인 이하이면 0, 50 – 100인이면 1, 100인 이상이면 2인 범주형 변수로 가변수 $z_1$과 $z_2$를 이용하여 분석

| 범주 | $z_1$ | $z_2$ |
|---|---|---|
| 50인 이하 | 0 | 0 |
| 50-100인 | 1 | 0 |
| 100인 이상 | 0 | 1 |

# 다중 로지스틱 회귀 예제 II

- 로지스틱 회귀모형

$$\Pr(Y = 1|x) = p(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z_1 + \beta_4 z_2)$$

- 회귀계수 추정값

| 입력변수 | 회귀계수 | 표준오차 | $\chi^2$ | 유의확률 |
|---|---|---|---|---|
| 절편 | 12.221 | 3.594 | 11.562 | 0.000 |
| $x_1$ | -1.072 | 0.393 | 7.438 | 0.006 |
| $x_2$ | 11.524 | 4.131 | 7.781 | 0.005 |
| $z_1$ | 0.698 | 0.357 | 3.940 | 0.047 |
| $z_2$ | 2.614 | 0.791 | 10.907 | 0.001 |

# 다중 로지스틱 회귀 예제 III

- 해석
  - 모든 변수가 유의수준 $\alpha = 0.05$에 비해 작은 유의확률을 갖으므로 출력변수를 설명하는데 유의함
  - $\hat{\beta}_1 < 0$이므로 부채비율이 높을 수록 회사채의 신용이 낮아짐
  - $\hat{\beta}_2 > 0$이므로 현금회전율이 높을 수록 회사채의 신용이 높아짐
- $z_1$과 $z_2$는 가변수이므로 그 계수에 대한 해석에 주의
  - $z_1$: $50-100$인 규모의 회사와 50인 미만의 회사의 회사채 신용등급의 오즈비는 $\exp(0.698) = 2.00$배 정도 높음
  - $z_2$: 100인 이상 규모의 회사와 50인 미만의 회사의 회사채 신용등급의 오즈비는 $\exp(2.614) = 13.655$배 정도 높음
- 가변수인 $z_1$과 $z_2$의 해석은 항상 50인 미만의 회사(가변수 값이 모두 0인)를 기준으로 비교하였는데 가변수를 다르게 코딩하면 회귀계수의 추정치가 달라짐.
- 그러나 상대적인 오즈비는 항상 일정하고 그 결과의 해석에는 차이가 없음

# 로지스틱 모형을 이용한 분류 I

- 0과 1사이의 적당한 수 $c$를 절단값으로 선택하여,
  - $\Pr(Y = 1|x)$가 $c$보다 크면 자료를 $Y = 1$인 클래스로 분류하고,
  - $\Pr(Y = 1|x)$가 $c$보다 작으면 자료를 $Y = 0$인 클래스로 분류

# 로지스틱 모형을 이용한 분류 II

- 절단값 $c$의 결정시 고려사항
- 사전정보: 두번째 범주의 자료($y = 1$인 자료)가 많이 나타난다면 절단값을 작은 값으로 정할 수 있음.
- 손실함수: 두번째 범주의 자료를 잘못 분류하는 손실이 첫번째 범주의 자료를 잘못 분류하는 것에 비하여 손실 정도가 심각하게 큰 경우에는 절단값 $c$를 작게 잡을 수 있음

  (예) 스팸 이메일로 분류, 암진단
- 전문가 의견, 민감도, 특이도 등
- 가령 $\Pr(Y = 1|x) > c$이면 클래스 1로 분류하는 규칙을 생각해 보자. 이 규칙은 $\log\left(\Pr(Y = 1|x)/(1 - \Pr(Y = 1|x))\right) > \log(1 + c)/c$이면 클래스 1로 분류하는 것과 동일하다. 즉, $c^* = (1 + c)/c$로 놓으면 $\beta_0 + \beta_1 x > \log(c^*)$이면 클래스 1로 분류하는 것이다. 따라서 로지스틱 모형의 분류경계는 선형임을 알 수 있음.
- 분류경계가 비선형인 경우에는 다항 로지스틱 모형으로 추정하거나 의사결정나무나 신경망모형 등의 비선형 분류경계를 찾는 방법들을 고려