Linear/Logistic Regression Analysis

Computational Linguistics @ Seoul National University

Hyopil Shin

회귀분석

- 회귀분석(regression analysis) – 영국의 우생학자 골턴(Francis Galton)이 처음 사용한 것으로 아버지의 키와 아들의 키 사이의 관계가 양의 직선관계이나 기울기가 45도 보다 작아 아들의 키는 인간의 평균키로 회귀하려는 경향이 있다고 주장한 데서 비롯됨
- Pearson은 런던 지역에서 살고 있는 1,078 부자의 키에 관한 자료를 수집하여 아버지의 키별로 아들의 평균키를 구하고 이들 간의 관계를 나타내는 직선의 방정식 을 제시

$$\hat{Y} = 33.73 + 0.516X$$

- 회귀분석은 변수들간의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계적 모델

회귀분석(From『언어학과 통계모델』)

회귀분석(regression analysis)은 여러 변수들 사이의 관계를 알아보고자 하는 경우에 사용할 수 있는 분석 방법이다. 두 변수 사이의 관계를 알기 위해서는 두 변수의 상관계수를 구하는 방법도 있지만, 상관계수는 두 변수 사이에 어떤 함수관계가 있는지는 보여 주지 못한다. 그러나 변수들 사이의 관계를 규명한 후 이를 이용하여 한 변수의 값으로부터 다른 변수를 예측하는 경우에 그 관계를 함수식으로 나타낼 필요가 있다. 대학 수학능력시험에서 높은 성적을 받은 학생이 낮은 성적을 받은 학생보다 대학에서 좋은 성적을 낸다면 수학능력시험에서 원점수 370점을 받은 학생의 학점이 얼마 정도 될 것이라고 예측할 수 있을까? 그리고 이 예측에 대한 오차의 범위는 어느 정도라고 할 수 있는가? 등이 그 예다.

회귀분석의 창시자 골턴(Francis Galton)은, 아버지와 아들의 키를 각각 x_i, y_i 라 하고 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ 과 같이 n 쌍의 부자 사이의 키에 대한 자료가 관측되었다고 하면 아버지와 아들의 키 사이에 다음과 같은 관계가 성립한다고 생각하였다.

$$y_i = M + \beta(x_i - M) + \epsilon_i, \quad i = 1, \cdots, n \tag{3.28}$$

회귀분석(From『언어학과 통계모델』)

여기서 M 은 아버지 키의 중앙값이며 ϵ_i 는 오차항이다. 골턴은 β 의 값이 0과 1사이에 있는 것으로 보고 β 값이 0보다 크므로 아버지의 키가 크면 아들의 키도 커지는 경향이 있으나 β 의 값이 1보다는 작으므로 아버지의 키가 아버지의 중앙값 M 보다는 아주 크더라도 아들의 키와 M 과의 차이는 아버지의 키와 M 과의 차이보다 작아지는 경향이 있다고 생각했다. 따라서 인간의 키는 그것의 중앙값으로 회귀(regression)한다고 결론을 내렸다. 이러한 종류의 분석을 회귀분석이라고 한다.

그러면 회귀에 따른 예측이 어떻게 이루어지는지 살펴보자. 그림 3.4

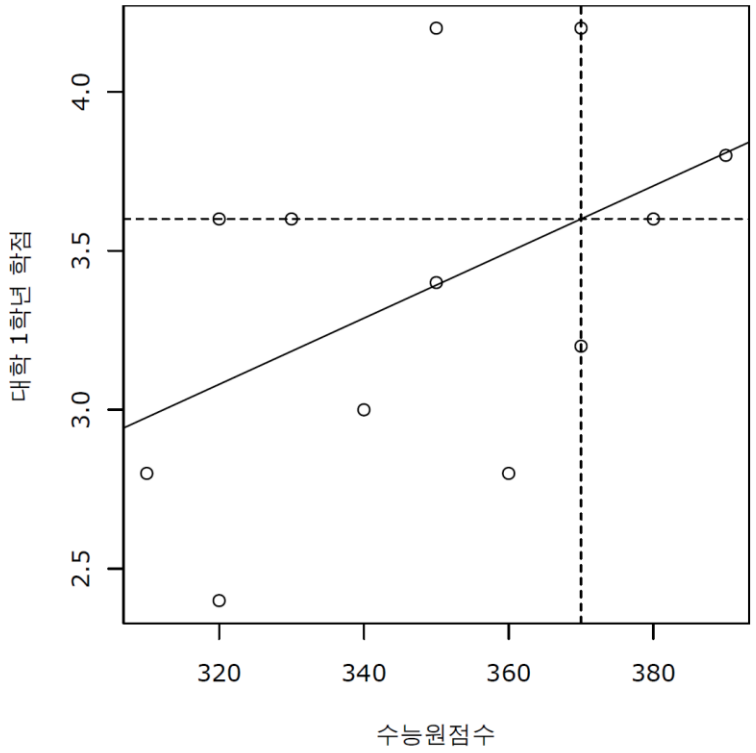


그림 3.4 수능점수에서 예측된 대학 1학년 학점

는 어느 대학 1학년 학생들의 수학능력 원점수(X 축)와 대학 학점(Y 축) 사이의 관계를 보여 주는 산포도다.

회귀분석(From『언어학과 통계모델』)

수능점수 370 점에 대해 예측된 학점 3.58 을 위해 직선을 사용하였다. 이 직선으로 수능점수에서 예측되는 학점을 구할 수 있다. 이렇게 수능점수와 학점 간의 경향성을 직선으로 표현할 수 있으면 최적의 직선을 구할 수 있고 수능점수에서 대학의 학점을 예측할 수 있다. 이러한 직선을 회귀선(regression line) 이라 한다. 예측은 어떤 X 값에 상응하는 Y 값을 찾음으로 이루어진다.

만일 값들이 정확히 직선 위에 위치한다면, 예측에 오차가 없을 것이다. 그러나 상관계수가 작을 경우 예측오차(prediction error) 가 생기게 된다. 따라서 회귀분석에서는 한 변수에서 다른 변수를 예측하고 예측한계를 결정하는 것이 중요하다. 이제 이 문제를 살펴보자.

회귀분석(From『언어학과 통계모델』)

3.2.1 최적선

회귀선이 최적선이 되는지를 알아보는 방법에 대해 생각해 보자. 예를 들어, 그림 3.4 에서 A 와 B 두 학생이 동일한 수능점수 370 점을 받았는데 이로 예측될 수 있는 학점은 3.58 이다. 그러나 실제 받은 학점은 각각 3.78 과 2.82 라고 하자. 예측된 값을 Y', 변수의 실제 값을 Y 라 하고 두 학생의 예측 오차를 구하면 다음과 같다.

$$\text{A–오차}: (Y - Y') = 3.78 - 3.58 = 0.20$$
$$\text{B–오차}: (Y - Y') = 2.82 - 3.58 = -0.76$$

회귀선은 전체적으로 산포도에 대한 예측오차$(Y - Y')$ 가 최소화되는 방법으로 그려진다. 따라서 최적선에서는 모든 예의 제곱된 예측오차의 합이 가능한 작아져야 하며, 적절한 회귀선이 되려면 다른 직선에 비해 오차 제곱합이 작아야 한다. 이를 최소제곱기준(least-squares criterion)이라 한다.

수능점수로부터 학점을 예측하는 것과 같이 X 에서 Y 를 예측할 때 X 는 독립변수, Y 는 종속변수라고 한다. X 에서 Y 를 예측하는 것은 X 에 Y 를 회귀시키는 것이 된다. 이제 이 예측이 어떻게 이루어지는지 알아보자

회귀분석(From『언어학과 통계모델』)

3.2.2 회귀식

산포도를 바탕으로 한 회귀선은 회귀식(regression equation)에 의해 결정된다. 이는 회귀선을 방정식으로 표시한 것으로 기울기와 절편으로 이루어진다. 기울기는 회귀선의 각과 방향을 나타내며 절편은 $x = 0$ 일 때의 y의 값이다.

$$\hat{y} = a + bx \tag{3.29}$$

일반적으로 회귀식을 구할 때는 가우스(Gauss)의 최소제곱법이 널리 사용된다. 그림 3.5를 살펴보자.

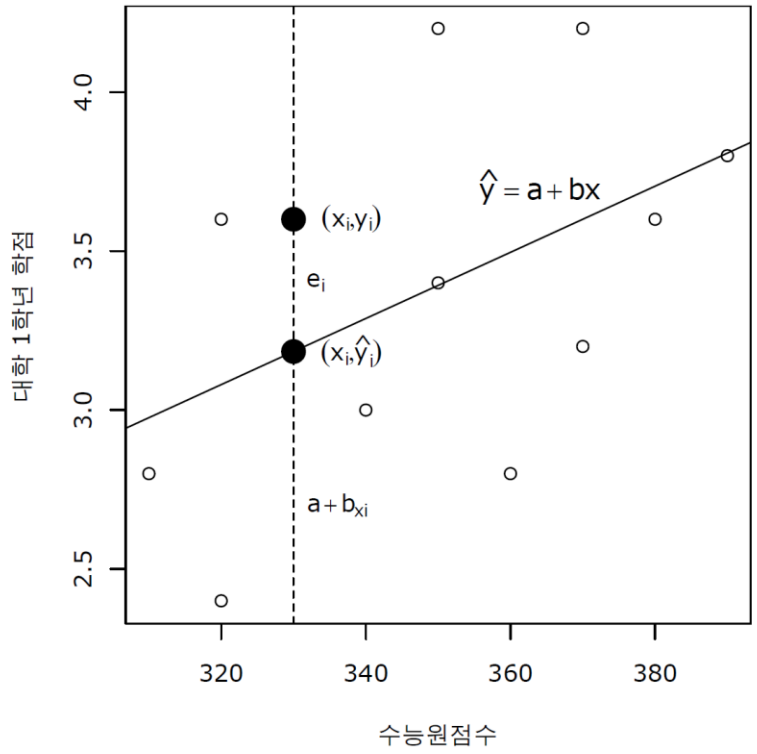


그림 3.5 수능점수와 학점의 관계

회귀분석(From『언어학과 통계모델』)

앞서 살펴본 대로 회귀선이 적절하기 위해서는 실제값과 추정값인 직선 상의 $\hat{y}$ 값 차이를 나타내는 잔차(residual error) e_i 가 가장 작을 때의 직선을 구해야 한다. 잔차는 다음과 같이 구해진다.

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - (a + bx_i) \end{aligned} \tag{3.30}$$

관측치가 n 개일 때 이를 모두 반영하기 위해 잔차의 합을 구해서 최소가 되는 값을 구해야 한다. 그러나 잔차가 $+$, $-$ 로 나타나 서로 상쇄되어 그 합은 0 이 되어 버린다. 이를 해소하기 위해 잔차의 제곱합을 이용한다.

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \tag{3.31}$$

회귀분석(From『언어학과 통계모델』)

잔차의 제곱합 S 가 최소가 되는 a, b 를 구하면 된다. 이를 위해 먼저 S 를 미지 변수인 a, b 로 각각 편미분한 값을 0 으로 둔다(이창효 · 김종배 2005)

$$
\begin{aligned}
S &= \sum (y_i - a - bx_i)^2 \qquad (3.32)\\
\frac{\partial S}{\partial a} &= \sum 2(y_i - a - bx_i)(-1) = 0\\
&\Rightarrow \sum y_i = \sum a + b\sum x_i\\
&\Rightarrow \sum y_i = na + b\sum x_i\\
\frac{\partial S}{\partial b} &= \sum 2(y_i - a - bx_i)(-x_i) = 0\\
&\Rightarrow \sum (y_i - a - bx_i)(x_i) = 0\\
&\Rightarrow \sum x_i y_i = a\sum x_i + b\sum x_i^2
\end{aligned}
$$

위 (3.32) 를 정규방정식 (normal equation) 이라 한다. 또한 회귀계수 a, b 는 이 정규방정식으로 구성된 연립방정식의 해로 다음과 같이 구할 수 있다.

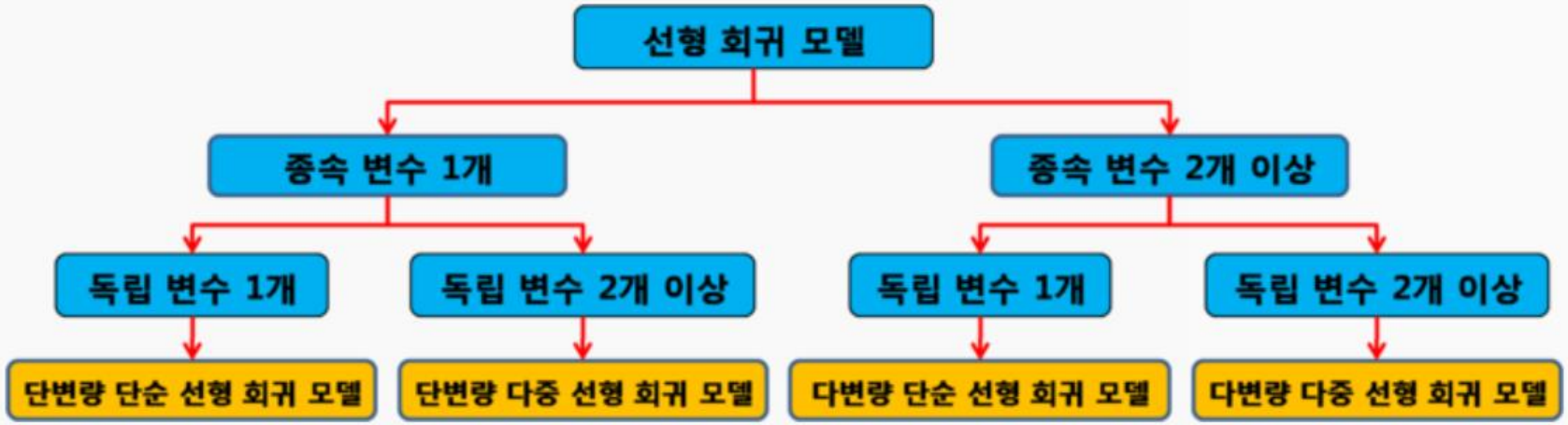
$$
\begin{aligned}
a &= \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} \qquad (3.33a)\\
&= \bar{y} - b\bar{x}
\end{aligned}
$$

$$
\begin{aligned}
b &= \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \qquad (3.33b)\\
&= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\end{aligned}
$$

선형회귀모형

- 어떤 변수에 다른 변수들이 주는 영향력을 선형적으로 분석하는 대표적인 방법이 선형회귀분석
- 선형회귀분석을 위해서는 선형회귀모델(linear regression model)을 만들어야 하는데, 여기서 모델은 수학 식으로 표현되는 함수를 의미하며, 영향을 주는 변수와 영향을 받는 변수로 구성됨
- 영향을 주는 변수는 독립변수 (independent variable) 또는 설명변수(explanatory variable) 등으로 불리며, 영향을 받는 변수는 종속 변수(dependent variable) 또는 반응변수(responsive variable)이라고 불림

선형회귀 모델



단변량 단순 선형 회귀 모델: univariate simple linear regression model

단변량 다중 선형 회귀 모델: univariate multiple linear regression model

다변량 단순 선형 회귀 모델: multivariate simple linear regression model

다변량 다중 선형 회귀 모델: multivariate multiple linear regression model

선형회귀모델

- 독립변수와 종속변수간 영향 관계
  - 최근 5년간 통화량, 환율, 실업률, 인구증가율이 물가에 미치는 영향을 알아 볼 때, 주 관심사는 물가이기 때문에 종속변수는 '물가지수'가 되고 통화량, 환율, 실업률, 인구증가율이 독립변수
  - 이 네가지 독립변수가 각각 물가지수에 미치는 영향은 회귀 분석을 통해 추정되는 계수(회귀 계수, regression coefficient)의 크기 및 방향성(+, -)으로 알 수 있음.
  - 계수의 크기가 0에 가까운 독립변수는 물가에 주는 영향력이 없고, 0보다 큰 양의 수를 가질 경우 해당 독립변수가 증가할 수록 물가지수도 올라감

단변량 단순선형회귀모델(univariate simple linear regression model)

- X라는 독립변수가 Y라는 종속 변수에 영향력을 주는 식
- X의 영향력은 β라는 계수의 크기와 부호로 표시됨
- α는 값이 변해도 Y의 변동에는 영향을 주지 않는 계수
- ε는 오차항(error term)이라 함

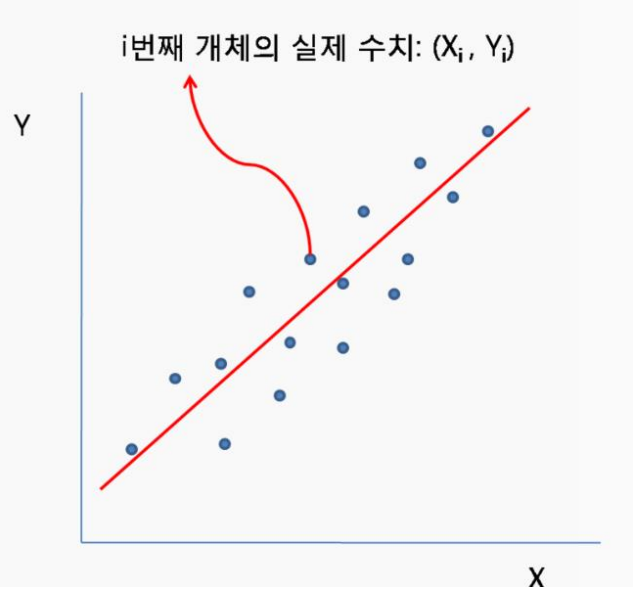
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

$$i = 1,2,\ldots,n$$

$$n : \text{개체수}$$

단순선형회귀모델

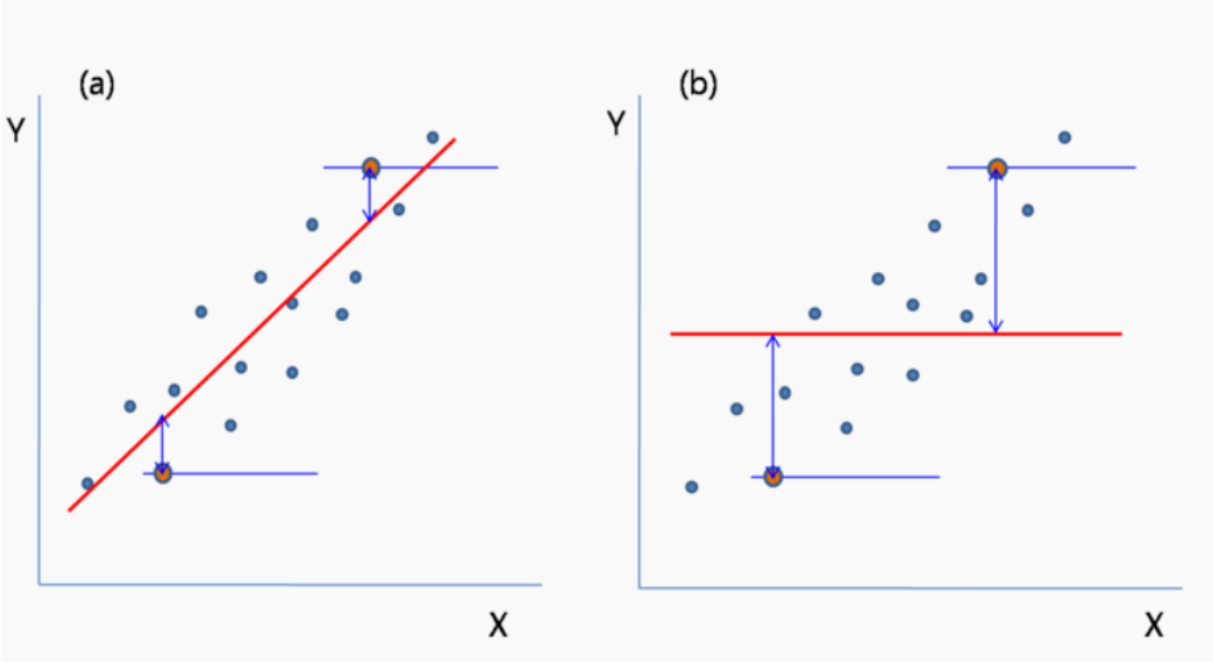
- 점들이 왼쪽에서 오른쪽 위로 올라가는 형태를 보임
- 붉은 직선은 2차원 좌표에 배치된 상태, 즉 변수 x와 변수 Y 의 관계를 가장 잘 나타내는 가장 단순한 형태의 함수



$$Y = a + bX$$

단순선형회귀모델

- X와 y의 관계를 가장 잘 나타내는 것?
  - (X, Y)의 좌표로 나타낸 점들에 가장 가까이 있는 직선을 찾는다는 것
  - 모든 점들과의 거리의 합이 최소가 되는 직선을 찾는다는 것



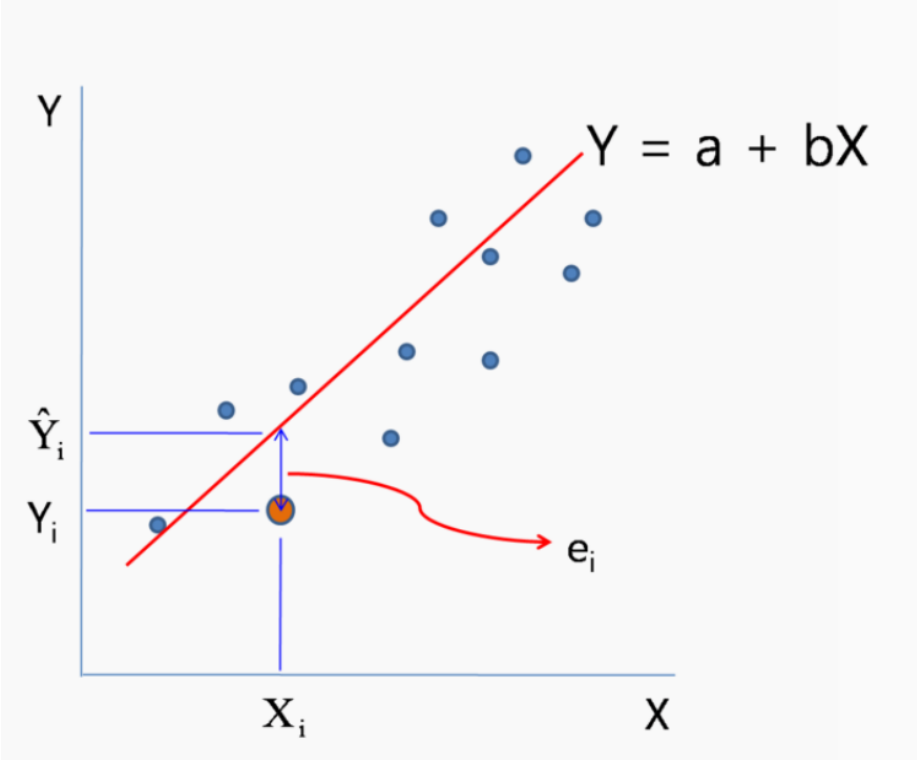
단순선형회귀모델

- e_i – 점들과 직선과의 거리
- SSE(sum of squared errors of prediction)를 최소화하는 a, b를 계산하면 X, Y 관계를 가장 잘 나타내 주는 직선, 선형회귀식이 나옴

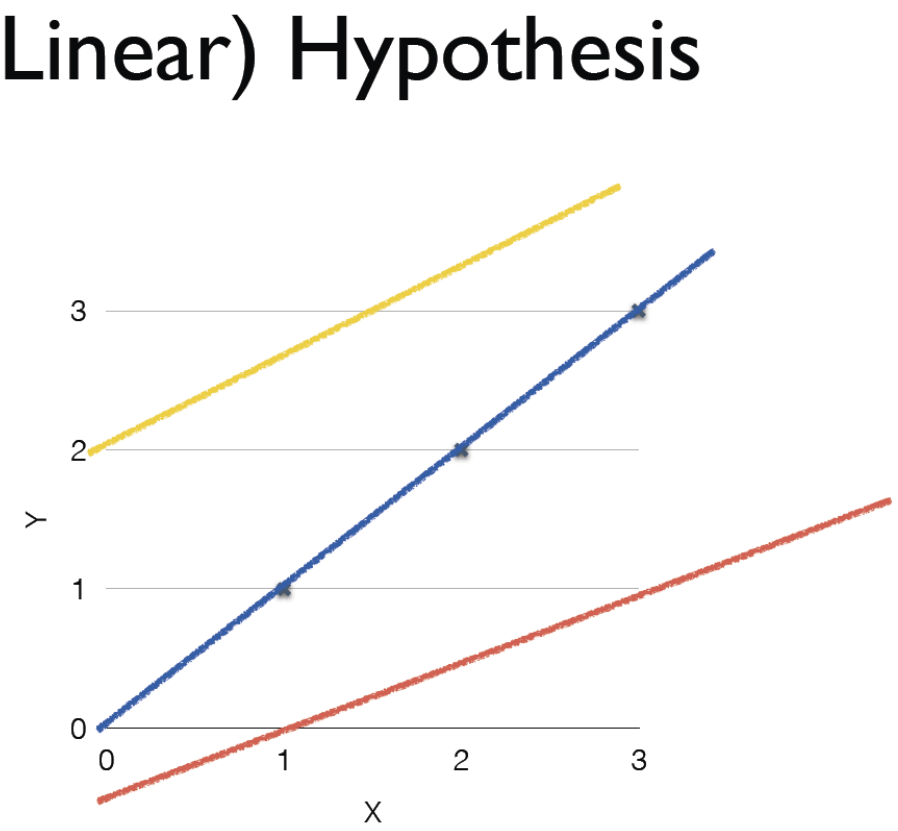
$$e_i = | Y_i - \hat{Y}_i | = | Y_i - a - bX_i |$$

$$SSE = \sum_{i=1}^{n} e_i^2$$

n : 개체수



Linear Hypothesis

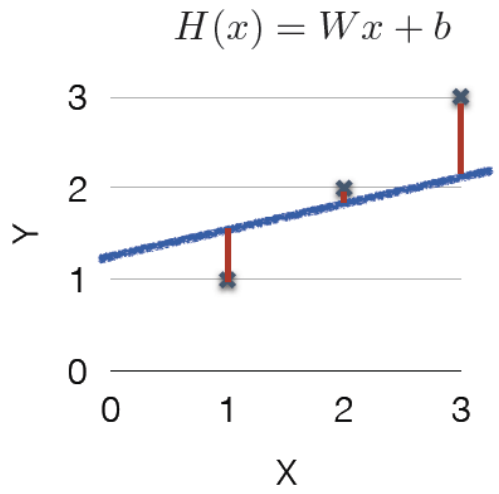


Cost function

- How fit the line to our (training) data

$$\frac{(H(x^{(1)}) - y^{(1)})^2 + (H(x^{(2)}) - y^{(2)})^2 + (H(x^{(3)}) - y^{(3)})^2}{3}$$

$$cost = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$$



회귀분석(

표 3.12 어느 백화점의 월간 광고비와 월간 판매량

월	1	2	3	4	5	6	7	8	9	10
월간 광고비	2	4	5	4	6	7	8	9	12	13
월간 판매량	45	53	50	45	55	60	78	90	112	121

(3.33)에 따라 기울기 b와 절편 a를 구하면 다음과 같다.

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} = \frac{709 \times 604 - 70 \times 5833}{10 \times 604 - 4900} \approx 17.47$$

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10 \times 5833 - 70 \times 709}{10 \times 604 - 4900} \approx 7.63$$

회귀선은 다음과 같다.

$$\hat{y} = 17.47 + 7.63x$$

$$\hat{y} = 17.47 + 7.63 \times 13 = 116.66$$

회귀분석(From『언어학과 통계모델』)

여기서 기울기와 절편은 가중치로 월간 광고비와 같은 자질을 반응값(월간 판매량)으로 사상시킨다. 이 가중치를 사용하여 선형함수(linear function)로 나타내면 다음과 같다.

$$\text{월간 판매량} = w_0 + w_1 \times \text{월간 광고비} \tag{5.34}$$

설명변수가 두 개 이상인 다중회귀분석의 경우, 월간 판매량에 영향을 미치는 요소로 월간 광고비 외에, 매달 쿠폰 발행 금액, 세일 일수와 같은 다른 요인들을 설정할 수 있다. 이를 변수로 설정하고 각 변수의 가중치를 결정하면 다음과 같은 선형함수가 가능하다.

$$\begin{aligned}\text{월간 판매량} = w_0 &+ w_1 \times \text{월간 광고비} \\ &+ w_2 \times \text{쿠폰 발행 금액} \\ &+ w_3 \times \text{세일 일수}\end{aligned} \tag{5.35}$$

회귀분석(From『언어학과 통계모델』)

이제 우리가 관찰하고자 하는 대상(가령 매출)은 자질 벡터로 표시할 수 있다. 예를 들어 어느 달의 매출에 대해 그 달의 월간 광고비가 5 이고, 쿠폰 발행 금액이 3 이고, 세일 일수가 15 일이라면 매출의 자질 벡터는 $\vec{f} = (5, 3, 15)$ 이 된다. 만일 이 경우에 가중치로 $\vec{w} = (w_0, w_1, w_2, w_3) = (2, -5, -3, 2.5)$ 의 벡터가 주어진다면, 예측되는 매출은 자질과 그 가중치를 곱하여 결정할 수 있다.

$$\text{매출} = w_0 + \sum_{i=1}^{N} w_i \times f_i \tag{5.36}$$

일반적으로 계산을 단순화하기 위해 w_0 에 대해 그 값이 1 인 자질 f_0 를 설정하고, y 의 값을 추정하기 위한 선형회귀를 다음과 같이 규정한다.

$$y = \sum_{i=0}^{N} w_i \times f_i \tag{5.37}$$

자질과 가중치는 벡터로 되어 있기 때문에 이 선형회귀는 벡터의 각각의 값을 곱한 후 다시 더하는 과정이다. 이는 내적(dot product)이 된다. 두 벡터 사이의 내적 $a \cdot b$ 는 다음과 같이 정의된다.

$$a \cdot b = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n \tag{5.38}$$

이는 자질 벡터와 가중치 벡터 사이의 내적과 마찬가지다.

$$y = w \cdot f \tag{5.39}$$

Logistic Regression(From『언어학과 통계모델』)

5.8.2 로지스틱 회귀

선형회귀는 설명변수에 따른 반응 변수의 대응 값을 예측하는 데 유용하다. 한편 반응 변수가 실수 값이 아닌 이진형 (binary) 으로 취해지는 경우가 있다. 예를 들어 어떤 관찰 대상이 참 (true) 인지 아니면 거짓 (false) 인지의 이진 분류만이 필요한 경우, 선형회귀에서처럼 실수인 값보다는 어떤 자질에 대해 참 (1), 거짓 (0) 을 나타내는 분류자(classifier)를 생각해 볼 수 있다. 더 나아가 그 값으로 1, 0 을 도출하는 것보다 어떤 대상이 참인 부류나 거짓인 부류에 속할 확률을 제시할 수 있는 모형이 필요하다. 이렇게 값이 이진형으로 취해지는 회귀를 로지스틱 회귀 (logistic regression) 라 한다.

선형회귀 모델을 이용하여 이런 이진 결과를 도출할 수 있는 방법을 생각해 보자.

$$P(y = \text{true}|x) = \sum_{i=0}^{N} w_i \times f_i = w \cdot f \qquad (5.40)$$

이 수식은 결과 값 (y) 이 참으로 규정한 부류 (class) 에 속하면 관찰 대상 (x) 을 $y = 1$ 로, 그렇지 않은 경우는 $y = 0$ 으로 학습 (training) 할 수 있게 한다. 관찰 대상 x 는 자질 (f) 로 되어 있고 이 자질은 가중치 (w) 를 갖는데, 그 관찰 대상이 해당 부류에 속하는 경우인 참 (1), 그렇지 않은 경우인 거짓 (0) 으로 예측될 경우, 그 오차를 최소화할 수 있도록 가중치를 설정해야 한다. 이렇게 가중치가 설정되고 나면 관찰 대상에 대한 자질과 그 가중치의 내적에 의하여 참, 거짓 부류의 확률을 구할 수 있다.

그러나 이 수식은 확률이 0 과 1 사이의 값이어야 하는 확률의 기본 정의를 만족시키지 못한다. 왜냐하면 자질과 가중치의 내적에 의해 도

Logistic Regression(From『언어학과 통계모델』)

출되는 값은 $-\infty$ 와 ∞ 사이의 값이기 때문이다. 따라서 0 과 1 사이의 값을 갖는 이진 분류의 확률이 아니라 두 확률의 비율(ratio) 로, 즉 한 부류에 속할 확률과 그렇지 않을 확률 비로 결과를 예측하게 된다. 이 비율은 통계학에서 승산(odds) 이라 불리는 것으로 $\frac{p}{1-p}$ 로 구해진다. 이 승산비는 0 보다 크고 무한대보다 작은 값으로 나타나며, 확률값이 0 에 가까우면 작은 값으로, 1 에 가까우면 큰 값으로 나타난다. 예를 들어 어떤 사건이 일어날 확률이 0.8 이고 일어나지 않을 확률이 0.2 라면 일어날 사건의 승산비(odds ratio) 는 $\frac{0.8}{0.2} = 4$ 이다. 이제 이 선형 모형에서 결과 y 가 참일 승산은 다음과 같이 구해진다.

$$\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)} = w \cdot f \tag{5.41}$$

승산비는 0 과 무한대 사이의 값으로 나타나기 때문에 이 수식의 좌변과 우변은 같지 않다. 즉, 좌변은 0 과 무한대 우변은 $-\infty$ 와 ∞ 사이의 값으로 나타나기 때문에 좌변에 자연로그를 붙여 양쪽이 다 $-\infty$ 와 ∞ 사이의 값을 취하도록 해야 한다.

$$\ln\left(\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)}\right) = w \cdot f \tag{5.42}$$

승산의 로그를 취한 것을 로짓 함수(logit function) 라 한다.

$$\text{logit}(P(x)) = \ln\left(\frac{P(x)}{1 - P(x)}\right) \tag{5.43}$$

$P(y = \text{true})$ 를 구하기 위해 수식 (5.42) 를 전개해 보자.

$$\ln\left(\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)}\right) = w \cdot f$$

$$\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)} = e^{w \cdot f}$$

$$P(y = \text{true}|x) = (1 - P(y = \text{true}|x))e^{w \cdot f}$$

$$P(y = \text{true}|x) = e^{w \cdot f} - P(y = \text{true}|x)e^{w \cdot f}$$

Logistic Regression(From『언어학과 통계모델』)

$$P(y = \text{true}|x) + P(y = \text{true}|x)e^{w \cdot f} = e^{w \cdot f}$$

$$P(y = \text{true}|x)(1 + e^{w \cdot f}) = e^{w \cdot f}$$

$$P(y = \text{true}|x) = \frac{e^{w \cdot f}}{1 + e^{w \cdot f}} \tag{5.44}$$

$P(y = \text{true}|x)$ 가 이렇게 구해지면 두 값의 합은 1 이어야 하기 때문에 $P(y = \text{false}|x)$ 는 다음과 같이 구해질 수 있다.

$$P(y = \text{false}|x) = \frac{1}{1 + e^{w \cdot f}} \tag{5.45}$$

일반적인 로짓 함수는 (5.46) 의 형태로 되어 있기 때문에 $P(y = \text{true}|x)$ 수식의 분자, 분모에 $e^{-w \cdot f}$ 를 곱하면 로지스틱 회귀를 이룰 수 있는 (5.47) 과 같은 로짓 함수의 형태가 된다.

$$\frac{1}{1 + e^{-x}} \tag{5.46}$$

$$P(y = \text{true}|x) = \frac{e^{w \cdot f}}{1 + e^{w \cdot f}} = \frac{1}{1 + e^{-w \cdot f}} \tag{5.47}$$

(5.47) 과 마찬가지로 합이 1 인 확률의 성질을 이용하여 $P(y = \text{false}|x)$ 는 다음과 같이 표시할 수 있다.

$$P(y = \text{false}|x) = \frac{e^{-w \cdot f}}{1 + e^{-w \cdot f}} \tag{5.48}$$

이제 로지스틱 회귀를 이용하여 관찰 대상이 참, 거짓의 이진법에 따라 어떻게 분류되는지 살펴보자. 확률값이 더 큰 대상이 참인 부류에 속한다. 따라서 다음의 경우를 참이라고 할 수 있다.

$$P(y = \text{true}|x) > P(y = \text{false}|x)$$

Logistic Regression(From『언어학과 통계모델』)

$$\frac{P(y=\text{true}|x)}{P(y=\text{false}|x)} > 1$$

$$\frac{P(y=\text{true}|x)}{1-P(y=\text{true}|x)} > 1$$

$$e^{w \cdot f} > 1 \qquad ((5.42)\text{에 따라 승산비로 치환})$$

$$w \cdot f > 0 \qquad \text{지수를 없애면}$$

$$\sum_{i=0}^{N} w_i f_i > 0 \qquad (5.49)$$