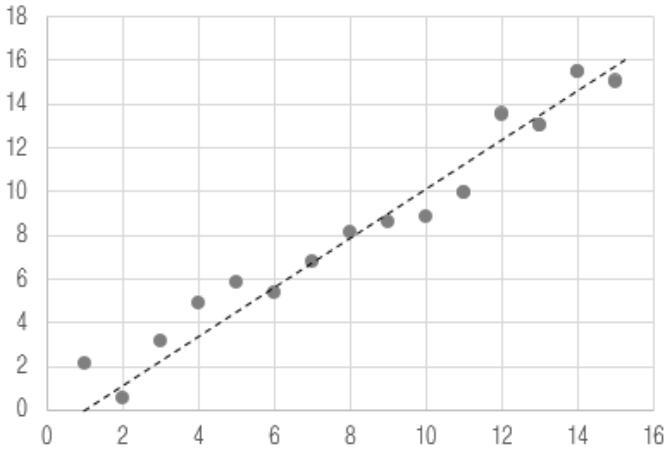
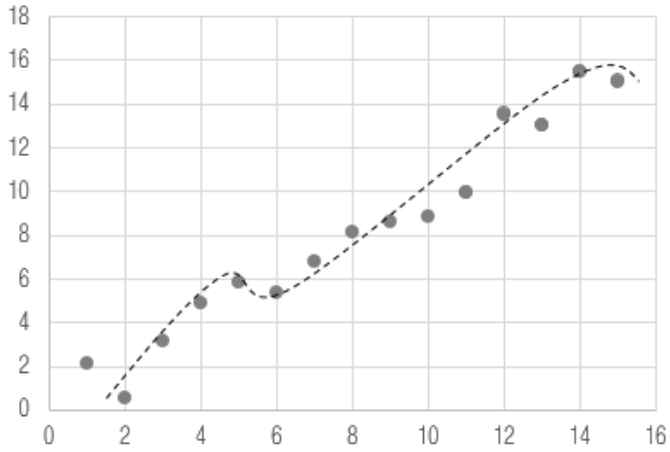
Linear Regression

회귀분석이란

- **입력 변수인 X의 정보를 활용하여 출력 변수인 Y를 예측하는 방법**
- 회귀분석 중 간단한 방법으로는 선형회귀분석(좌측 그림)이 있으며, 이를 바탕으로 더 복잡한 회귀분석(우측 그림)이 개발

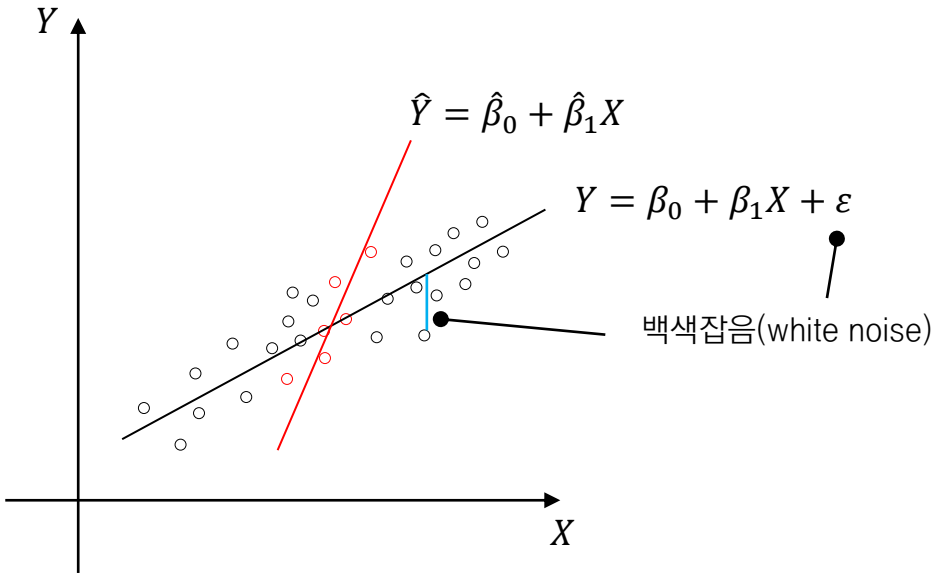




단순선형회귀의 회귀식

- 입력 변수가 X, 출력 변수가 Y일 때, 단순선형회귀의 회귀식은 검은선으로 나타낼 수 있음
- β_0는 절편(intercept), β_1은 기울기(slope)이며 합쳐서 회귀계수(coefficients)로도 불림

- 검은 점: 모집단의 모든 데이터
- 빨간 점: 학습집합의 데이터
- **실제 β_0와 β_1은 구할 수 없는 계수로 데이터(학습집합)를 통해 이 둘을 추정**해서 사용
- 모집단의 특징을 잘 설명할 수 있는 학습집합을 선택하는 것이 중요



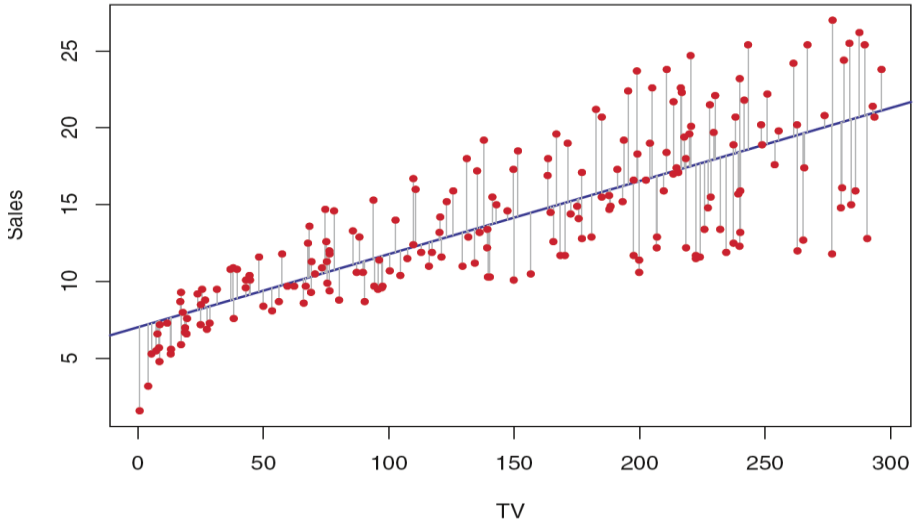
잔차(residual)의 의미

- 회귀계수의 추정에 대해 알아보기 전에 잔차의 의미를 알아야 함
- 잔차는 다음 식과 같이 정의되며, **실제 출력 변수와 예측한 출력 변수의 차**를 나타냄

$$e_i = y_i - \hat{y}_i$$

- 잔차를 그림으로 나타내면 오른쪽 그림과 같음
- 잔차의 제곱합(RSS; Residual Sum of Squares)는 아래와 같이 표현 가능

$$RSS = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

회귀계수의 추정

- 회귀계수는 RSS를 최소화 하는 방향으로 추정

$$RSS = \sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2$$
$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

- x_i와 y_i는 주어진 학습집합의 데이터이므로 위 식에서 변수는 $\hat{\beta}_0$과 $\hat{\beta}_1$임
- 따라서, RSS를 최소화하기 위해 위 식의 $\hat{\beta}_0$과 $\hat{\beta}_1$로 편미분을 하면 아래의 식이 도출(이 방법을 Least Square Method라고 부름)
- $\hat{\beta}_1$을 먼저 구한 후, 이를 이용하여 $\hat{\beta}_0$를 계산

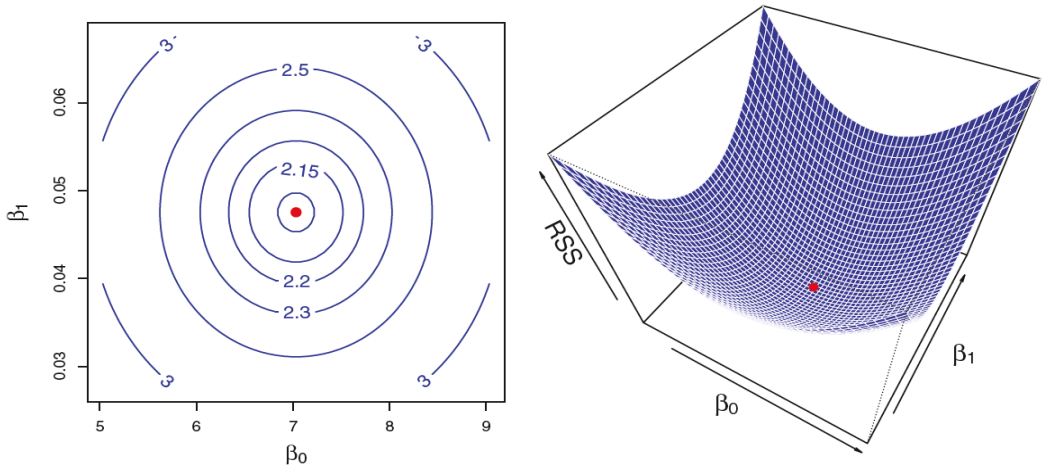
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$where\ \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ ,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

회귀계수의 추정

- 회귀계수의 추정을 그림으로 표현하면 아래와 같음
- 왼쪽 그림은 x축이 $\hat{\beta}_0$, y축이 $\hat{\beta}_1$일 때, RSS 등고선을 나타낸 그림
- 동그라미의 중심으로 갈수록 RSS가 점점 감소
- 이를 3차원으로 나타낸 그림이 오른쪽 그림
- **RSS가 가장 작은 지점의 $\hat{\beta}_0$과 $\hat{\beta}_1$를 찾는 것이 목표**



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

선형회귀의 정확도 평가

- 선형회귀의 정확도 평가는 크게 두 가지를 활용: RSE(Residual Standard Error)와 R^2
- RSE는 다른 말로 MSE(Mean Squared Error)이며 RSS를 표준화한 개념

$$RSE = MSE = \sqrt{\frac{1}{n-2}RSS}$$

- RSE는 출력 변수의 크기에 따라 값이 달라지는 성질이 있음
- 즉, **scale이 정해져 있지 않아 RSE는 객관적이 기준이 될 수 없다**

- RSS와 RSE는 일반적으로 우측의 표와 같은 표기법을 사용
- 하지만, Error라는 표현은 정확하지 않으며 Residual이 맞는 표현임

Source of Variation	Sum of Squares	Mean Square
Regression	SSR	SSR
Error	SSE(=RSS)	MSE(=RSE)
Total	SST(=TSS)	

선형회귀의 정확도 평가

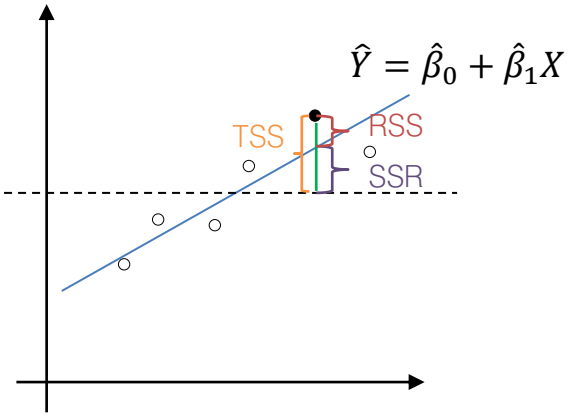
- R^2는 RSE의 단점을 보완한 평가지표로 0~1의 범위를 가짐
- R^2은 다음 식과 같이 표현

입력 변수로 설명할 수 없는 변동

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \qquad where\ TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- TSS는 Total Sum of Squares의 약자로 출력 변수의 변동을 의미
- R^2은 설명력으로 입력 변수인 X로 설명할 수 있는 Y의 변동을 의미
- R^2이 1에 가까울 수록 선형회귀 모형의 설명력이 높다는 것을 뜻함

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{RSS}}$$



단순선형회귀의 확장

- 단순선형회귀는 입력 변수의 종류가 하나일 때 사용이 가능
- 입력 변수가 여러 종류인 경우 단순선형회귀를 여러 번 사용함으로써 단순선형회귀의 확장이 가능
- 2강의 광고미디어 예에 위의 방법을 적용하면 아래 표와 같음
- 판매량에 영향을 미치는 라디오 광고와 신문 광고 각각에 대해 단순선형회귀를 수행

회귀 계수 / 회귀 계수의 표준 편차 / 회귀 계수의 유의성을 판단하는 통계치

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

- p-value: 유의수준으로 주로 0.05 이하의 값이면 해당 변수가 출력 변수에 영향을 미친다고 판단
- 라디오 광고 예산의 증가로 0.203의 판매량이 증가
- 신문 광고 예산의 증가로 0.055의 판매량이 증가
- **두 예산을 동시에 증가하였을 때 판매량을 예측하기 어려움**

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	< 0.0001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

단순선형회귀의 확장

- 입력 변수의 복합적인 변화에 따른 출력 변수의 변화를 예측하기 위한 모형이 필요
- 아래와 같이 단순선형회귀를 다중선형회귀로 확장

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\Downarrow$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

- p: 입력 변수의 종류
- 회귀계수인 $\hat{\beta}_i$를 구하는 방법은 단순선형회귀와 동일
- 입력 변수인 $X_1, X_2, \cdots, X_p$는 서로 독립임을 가정

단순선형회귀의 확장

- 광고미디어 예에 다중선형회귀를 적용하면 아래와 같음
- 신문 광고의 경우 단순선형회귀에서는 출력 변수인 매출과 연관이 있었지만, 다중선형회귀에서는 p-value가 0.86으로 높아 매출에 유의미한 영향을 미치지 못함
- **신문 광고의 단순선형회귀에서 TV 광고와 라디오 광고의 영향력을 무시했기 때문에 생긴 결과**

TV radio newspaper

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

단순선형회귀의 확장

- 신문 광고의 매출에 대한 영향력이 다중선형회귀에서 사라진 원인을 살펴보기 위해 상관관계 표를 제시

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- 라디오 광고와 신문 광고의 상관관계 계수(0~1의 범위)가 0.3541로 낮지 않음
- 이는 라디오 광고비가 더 투자되는 시장에서 신문광고비도 더 투자되는 경향을 나타냄
- 즉, 신문 광고가 매출에 영향이 없음에도 불구하고, 라디오 광고에 기인하여 매출이 증가하는 것처럼 보임

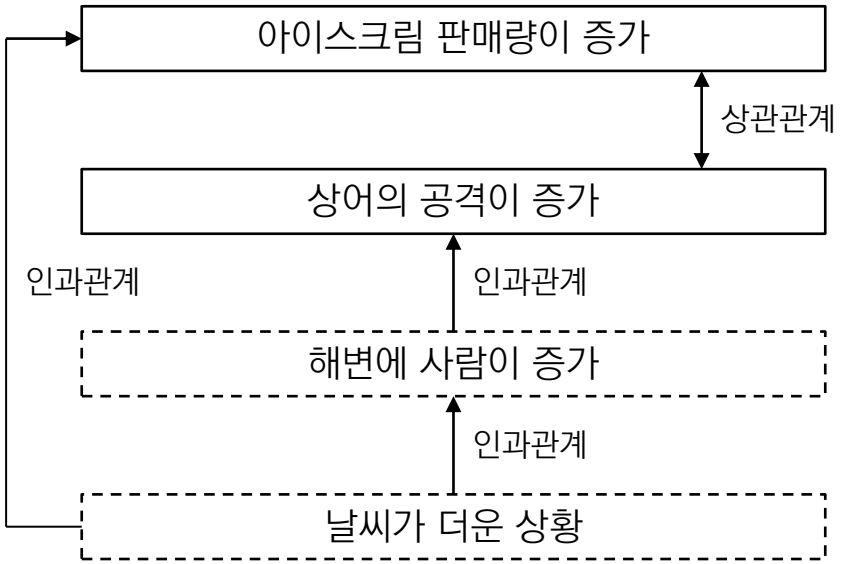
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

상관관계와 인과관계

- 데이터 마이닝에서 상관관계와 인과관계는 명확히 구분되어서 사용해야 하며 이 차이는 아래와 같음

◆ **상관관계(Correlation)**: 두 가지 변수간의 어떤 선형적 관계를 갖고 있는 지를 분석하는 통계적 기법

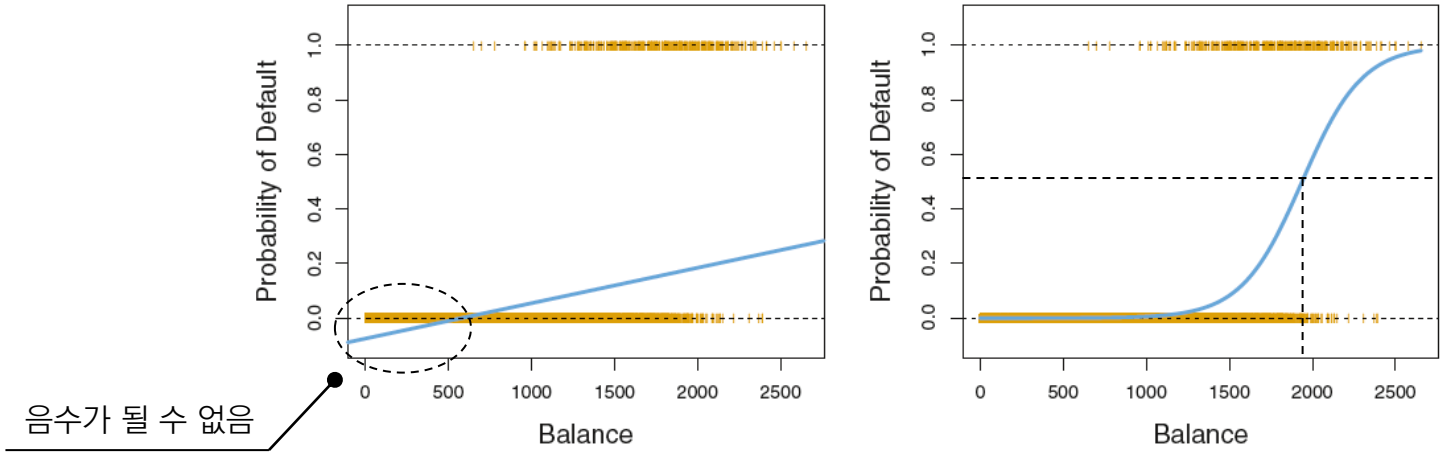
◆ **인과관계(Causality)**: 외부 요인을 통제한 상태에서, 두 변수가 원인과 결과의 관계를 맺고 있는 지를 분석



로지스틱회귀

로지스틱회귀란

- 로지스틱회귀는 출력 변수를 직접 예측하는 것이 아니라, 출력 변수가 1 or 0(binary)에 속할 확률을 모델링함
- 좌측 그림: 단순선형회귀를 이용했을 때, 잔고액(balance)에 따른 파산 확률(default)
- 우측 그림: 로지스틱회귀를 이용했을 때, 잔고액에 따른 파산 확률
- 약 잔고액이 2000 전후로 파산 확률을 결정할 수 있음



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

로지스틱 함수(logistic function)

- 로지스틱회귀의 식은 아래와 같음
- X는 입력변수, Y는 출력변수가 1이 될 확률

$$Y = p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

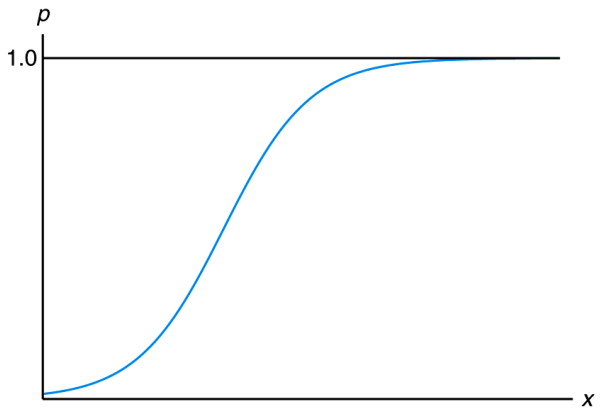
- β_1이 양수인 경우, 위의 함수는 X가 $-\infty$로 향할 때 $p(X)$가 0, X가 ∞로 향할 때 $p(X)$가 1로 수렴(그림 참조)

$$odds = \frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$

도박에서 승률을 의미

$$logit = \log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

회귀식을 선형으로 변환하는 함수



Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (1993). *Probability and statistics for engineers and scientists* (Vol. 5). New York: Macmillan.

로지스틱회귀

회귀계수의 추정

- 단순(다중)선형회귀의 least square method를 사용하는 것이 아닌 maximum likelihood를 사용
- Likelihood function은 아래와 같고, 이를 최대화하는 β_0, β_1를 추정

$$Maximize_{\beta_0,\beta_1} \quad l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} \left(1 - p(x_{i'})\right)$$

- 위를 실제 사례(앞의 그림)에 적용하면 아래의 표와 같은 결과가 도출
- $\hat{\beta}_0$과 $\hat{\beta}_1$ 모두 유의하였고, 잔고액이 1 증가할 때마다 logit이 0.0055 증가

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

로지스틱회귀

다중 로지스틱회귀

- 단순선형회귀와 마찬가지로 **로지스틱회귀도 입력 변수가 여러 종류일 때로 확장이 가능**
- 입력 변수가 하나일 때와 마찬가지로 maximum likelihood 방법을 이용하면 회귀계수의 추정이 가능

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- 파산 확률에 영향을 주는 요인이 수입(income)과 학생 여부(student[Yes])가 추가되었을 때 다중 로지스틱회귀를 적용하면 아래 표와 같은 결과가 도출
- 잔고액과 학생 여부가 유의한 입력 변수였으며, 학생이면 파산할 확률이 낮다는 결과가 도출

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

회귀계수를 축소하는 이유

- 영향력이 없는 입력 변수의 계수를 0에 가깝게 가져간다면, 모형에 포함되는 입력 변수의 수를 줄일 수 있음
- 입력 변수의 수를 줄이면 크게 세 가지 장점이 있음

① 잡음(noise)을 제거해 모형의 정확도를 개선

② 모형의 연산 속도가 빨라짐

③ 다중공선성의 문제를 제거해 모형의 해석 능력을 향상

많은 모형에서 입력 변수들끼리 독립임을 가정하지만, 입력 변수들끼리 상관관계를 가지는 경우

입력 변수가 나이, 잔고액, 생년인 경우 나이와 생년은 같은 의미를 갖기 때문에 둘 중 하나를 제거

계수축소법

계수축소법의 종류

- 계수축소법은 기본적으로 다중선형회귀와 유사
- 다중선형회귀에서 잔차를 최소화했다면, 계수축소법에서는 잔차와 회귀계수를 최소화
- 계수축소법에는 크게 두 가지의 방법이 있음: Ridge 회귀, Lasso
- 아래 식은 다중선형회귀의 RSS이며, 다중선형회귀에서는 RSS가 최소화되는 회귀계수를 추정

$$minimize\ RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

- 계수축소법에서는 위 식에 회귀계수를 축소하는 항을 추가

$$minimize\ RSS + f(\beta)$$

계수축소법

Ridge 회귀

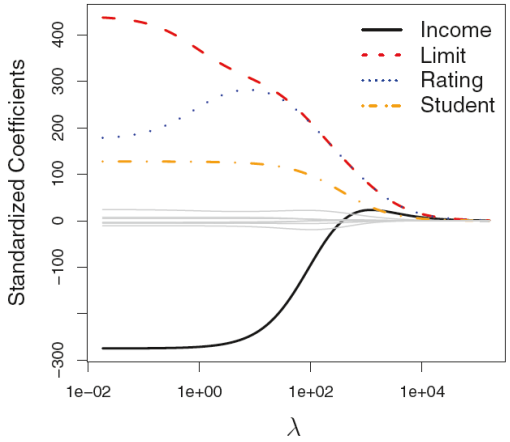
Ridge 회귀에서는 $f(\beta)$에 회귀계수의 제곱의 합을 대입

λ는 tuning parameter로 크면 클 수록 보다 많은 회귀계수를 0으로 수렴

회귀계수의 제곱의 합

$$minimize \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

- 오른쪽 그림은 λ가 커질수록 입력변수인 Income, Limit Rating, Student가 0으로 수렴하는 것을 표현
- 적절한 λ의 값은 데이터마다 달라지며, 현재는 e^4인 54.6의 값을 설정하였을 때 모든 입력 변수가 0으로 수렴함



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

계수축소법

Lasso

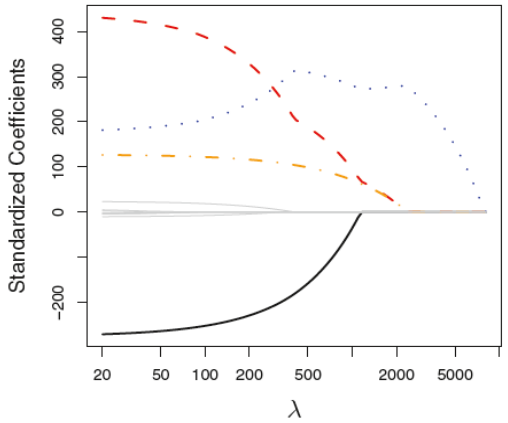
Lasso 회귀에서는 $f(\beta)$에 회귀계수의 절대값의 합을 대입

λ는 tuning parameter로 크면 클 수록 보다 많은 회귀계수를 0으로 수렴

회귀계수의 절대값의 합

$$minimize \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

- 오른쪽 그림은 λ가 커질수록 입력변수인 Income, Limit Rating, Student가 0으로 수렴하는 것을 표현
- 적절한 λ의 값은 데이터마다 달라지며, 5000이 넘는 값을 설정한 경우에 모든 입력 변수가 0이 됨



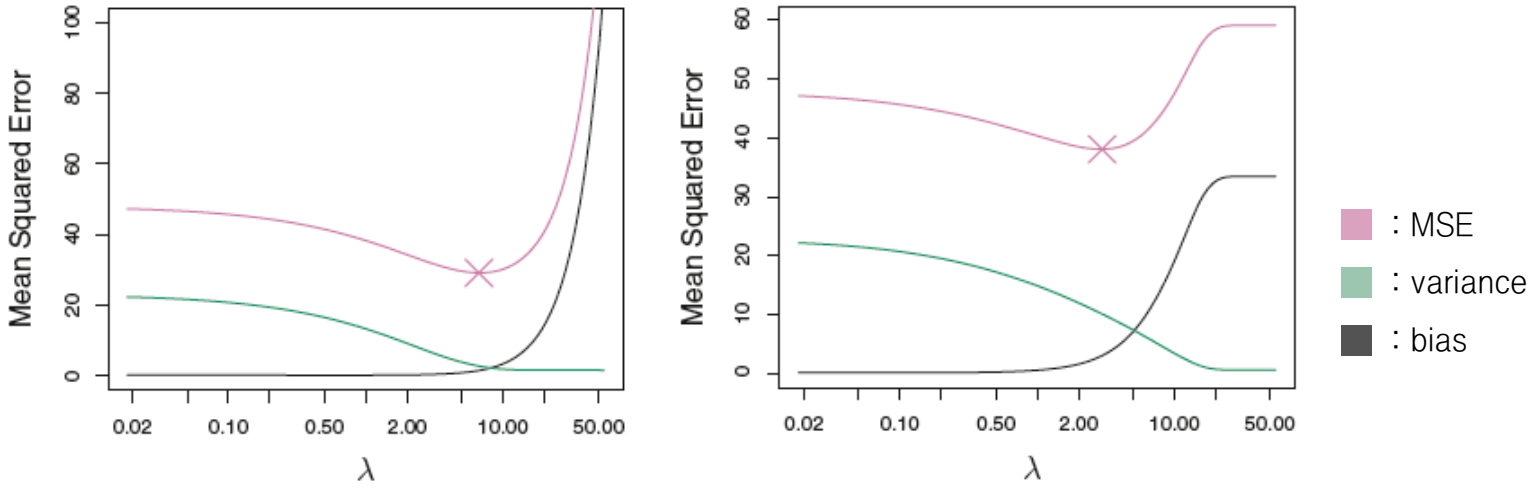
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

람다(lambda) 값의 설정

- 적절한 람다 값은 다음과 같은 방법으로 설정

> 람다 값을 변화시켜가며 MSE가 최소일 때의 람다를 탐색

- Ridge 회귀(좌측 그림)와 Lasso(우측 그림)의 람다에 따른 MSE의 변화는 아래 그림과 같음
- 한 모의 데이터에 적용하였으며 Ridge 회귀의 경우 8, Lasso의 경우 4 부근에서 MSE가 최소



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

계수축소법의 최적화 표현

- Ridge 회귀와 Lasso는 다음과 같은 최적화 방식으로 표현이 가능(상단: Ridge 회귀, 하단: Lasso)

$$minimize \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} {\beta_j}^2 \leq s$$

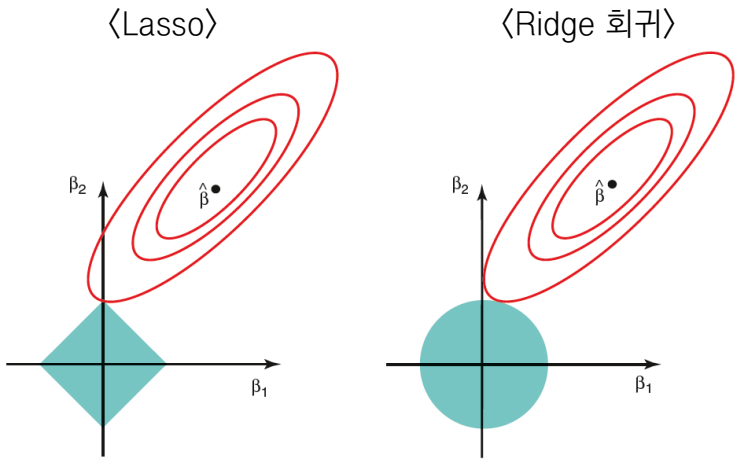
$$minimize \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

- λ 대신 s를 사용하여 회귀계수의 크기를 제한
- 위 식은 라그랑지안(Lagrangian) 최적화 기법으로 최적 해(최적 회귀계수)를 구할 수 있음

Ridge 회귀와 Lasso의 차이점

- Ridge 회귀와 Lasso의 가장 큰 차이점은 Ridge는 계수를 축소하되 0에 가까운 수로 축소하는 반면, Lasso는 계수를 완전히 0으로 축소함
- Ridge 회귀: 입력 변수들이 전반적으로 비슷한 수준으로 출력 변수에 영향을 미치는 경우에 사용
- Lasso: 출력 변수에 미치는 입력 변수의 영향력 편차가 큰 경우에 사용

- 초록색 그림: 회귀계수가 가질 수 있는 영역(feasible region)
- 빨간색 원: RSS가 같은 지점을 연결한 그림(가운데로 갈수록 오차가 작아짐)
- Ridge 회귀와 Lasso 모두 RSS를 희생하여 계수를 축소하는 방법
- Lasso의 경우 회귀계수가 0이 될 수 있지만, Ridge는 불가능



James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.