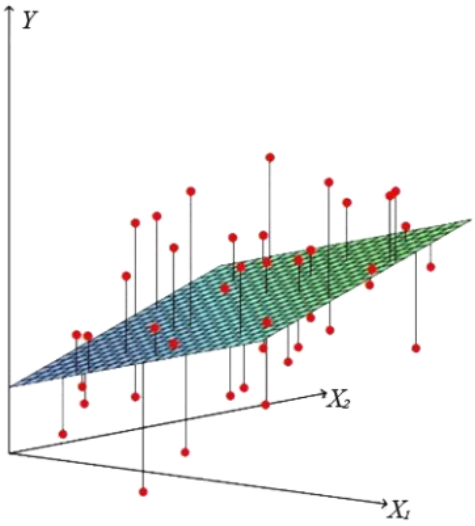
# 11. 다중회귀분석

- 다중회귀분석: 독립변수가 두개 이상인 경우
  - 여러개의 독립변수를 이용하면 종속변수의 변화를 더 잘 설명할 수 있음
  - 회귀 모형은 범주형 독립변수 **포함 가능**(> 더미변수를 이용)
- 다중선형회귀모형으로의 확장
  - 다중 선형회귀모형
    - Copyright by Multicampus Co., Ltd. All right reserved
    - 독립변수가 두개이상인선형회귀모형.
    - 여러개의독립변수를 이용하면종속변수의 변화에 대한 설명력 증가
    - 자료$((x_{1i}, x_{2i}, \cdots, x_{ki}), Y_i), i = 1, \cdots, n$ 에 다음의 관계식이 성립한다고 가정함.$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \qquad (i = 1, 2, \ldots, n)$

      $(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})$ = $E[Y_i]$ -기댓값
  - 다중 선형회귀모형
    - 오차항인 $\epsilon_1, \epsilon_2, ..., \epsilon_n$은 서로 독립인 확률 변수$\epsilon_i \sim N[0, \sigma^2]$ : 정규, 등분산, 독립
    - 회귀계수 α, $\beta_1, ..., \beta_k$, 와 $\sigma$2 은 미지인 모수로 상수임.
    - $\beta_j$ 의해석 : $x_j$ 를 제외한 나머지 모든 예측변수들을 **상수로 고정시킨 상태**에서
      $x_j$ 의 한 단위증가에 따른 $E[Y]$의 증분을 의미 (*j* = 1, … , *k*).
- 회계계수 $\alpha, \beta_1, ..., \beta_k$의 추정
  - 수직거리 제곱합
    - $SS(\alpha, \beta_1, ...\beta_k) = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_{1i} - \cdots - \beta_k x_{ki})^2$
      이 최소가 되도록 $\alpha$, $\beta_1, \ldots, \beta_k$ 를 추정.
    - 최소제곱 추정량 : $\widehat{\alpha}, \widehat{\beta_1}, ..., \widehat{\beta_k}$




- 다중회귀모형 분석 예시
  - 현재 영업중인 La Quinta Inn 호텔 중에서 랜덤하게 100곳의 영업자료를 수집
  - 다중회귀 모형의 설정
  - $Margin = \alpha + \beta_1 Number + \beta_2 Nearest + \beta_3 Office + \beta_4 College + \beta_5 Income + \beta_6 Disttwn + \epsilon$

| Margin | Number | Nearest | Office Space | Enrollment | Income | Distance |
|---|---|---|---|---|---|---|
| 55.5 | 3203 | 4.2 | 549 | 8 | 37 | 2.7 |
| 33.8 | 2810 | 2.8 | 496 | 17.5 | 35 | 14.4 |
| 49 | 2890 | 2.4 | 254 | 20 | 35 | 2.6 |
| 31.9 | 3422 | 3.3 | 434 | 15.5 | 38 | 12.1 |
| 57.4 | 2687 | 0.9 | 678 | 15.5 | 42 | 6.9 |
| 49 | 3759 | 2.9 | 635 | 19 | 33 | 10.8 |
| ... | ... | ... | ... | ... | ... | ... |

  - 다중선형회귀모형 추정결과
    - $margin = 38.14 - 0.0076 Number + 1.65 Nearest + 0.020 OfficeSpace + 0.21 Enrollment + 0.41 Income - 0.23 Distance$
    - $\widehat{\alpha} = 38.14$. 수학적인 해석은 모든 x값이 0일 때의 Profit Margin의 예상 값 (평균추정값)
      - 모든 x가 0인 게 관찰되지않았기때문에 해석하지 않음

- $\widehat{\beta_1} = -0.0076$. **다른변수는 고정되어 있고**, 반경3마일이내 호텔객실이 1개 증가하면,
  영업이익율은 평균0.0076% 감소.
- $\widehat{\beta_2} = 1.65$. **다른변수는 고정되어 있고**, 경쟁호텔과의 거리가 1마일 증가하면,
  영업이익율은 평균1.65% 증가.
- $\widehat{\beta_3} = 0.02$. **다른변수는 고정되어 있고**, 사무실 면적이 100평방피트 증가하면,
  영업이익율은 평균0.02% 증가.
- $\widehat{\beta_4} = 0.21$. **다른변수는 고정되어 있고**, 대학생거주자수가1000명 증가하면,
  영업이익율은 평균0.21% 증가.
- $\widehat{\beta_5} = 0.41$. **다른변수는 고정되어 있고**, 중간가구소득이 $1000 높은지역은,
  영업이익율이 평균 0.41% 증가.
- $\widehat{\beta_6} = -0.23$. **다른변수는 고정되어 있고**, 다운타운 중심으로부터 1마일멀어질수록,
  영업이익율은 평균0.23% 감소
- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 0.5251$
  영업이익율은 6개 설명변수에 의해 52.51% 설명됨.

- 범주형 독립변수가 포함된 회귀모형
  - 범주형 독립변수를 회귀모형에 포함하기 위해서는
    더미변수 (dummy variable) 기법을 사용.
  - 더미변수는 0 또는 1의 값을 갖는 변수로 아래와 같이 정의됨.

- 범주형 독립변수 예시 : 중고차 가격에 관한 예측 모형
  - 중고차 시장에서 차량의 주행거리와 색상이 차량의 가격에 어떤 영향을 미치는지를 파악하고자 2013년형 A브랜드 중고차 100대에 관한 자료를 수집
  - 종속변수 : Price(가격)
  - 독립변수 : Odometer(주행거리)
  - Color(차량색상, 범주형 : white/silver/other)
  - Color의 더미변수 : 범주의 수가 3개
    - 2개의 더미변수($D_1, D_2$)를 생성
    - $D_1 = \begin{cases} 1 & white\text{인경우} \\ 0 & white\text{가아닌경우} \end{cases}$
    - $D_2 = \begin{cases} 1 & silver\text{인경우} \\ 0 & silver\text{가아닌경우} \end{cases}$
    - Other인 경우 : $D_1 = 0 \ \& \ D_2 = 0$
  - $PRICE$ = 16701 − .0555($Odometer$) + 90.48($D_1$) + 295.48($D_2$)