3.2 다중선형회귀

- 두 개 이상의 설명변수에 각각 다른 단순선형회귀모델을 사용하는 것은 만족스럽지 않다.
  - 우선, 서로 다른 회귀방정식에 연관되어 있기에 예측방식이 명확하지 않다.
  - 각 회귀방정식은 다른 설명변수들을 고려하지 않는다.
- 대신 단순선형회귀모델을 확장하여 복수의 설명변수들을 직접 수용할 수 있는 것이 낫다.
  - 하나의 모델에서 각 설명변수에 다른 기울기 계수를 할당한다.
  - 서로 다른 설명변수가 p개 있다고 해보자.

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

  - β_j는 j번째 설명변수인 X_j와 반응변수 Y 사이의 연관성을 수량화하며, 다른 설명변수들을 변동되지 않ㄴ을 때 X_j의 한 유닛 증가가 Y에 미치는 평균 효과로 해석될 수 있다.

3.2.1 회귀계수의 추정

- 회귀계수 $\beta_0, \beta_1, \ldots, \beta_p$는 알려지지 않은 값이며 추정되어야 한다.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + + \hat{\beta}_p x_p$$

- $\beta_0, \beta_1, \ldots, \beta_p$는 잔차제곱합을 최소화하도록 선택된다.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_p x_{ip}\right)^2$$

- 단순회귀계수와 다중회귀계수는 다를 수 있다.
- 단순선형회귀에서는 상관관계가 유의하지만 다중회귀는 그 반대인 경우가 종종 발생하며, 이는 다른 설명변수들을 고려함으로써 기존의 단순선형회귀모델의 관계가 허위적이라는 것을 나타낸다.

3.2.2 몇 가지 중요한 질문

하나: 반응변수와 설명변수 사이에 상관관계가 있는가?

- p개 설명변수가 있는 다중회귀에서는 모든 회귀계수들이 0인지, 즉 $\beta_1 = \beta_2 = \ldots = \beta_p = 0$인지를 검사한다.
  - 귀무가설과 대립가설은 다음과 같다.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_a : \beta_j \neq 0 \ (for\ at\ least\ one\ j)$$

  - 이러한 가설검정은 F-통계량을 계산함으로써 이루어진다.

$$F^\star = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} \sim F(p, n-p-1)$$
$$where\ SSE = \sum (y_i - \bar{y})^2,\ SSR = \sum (\hat{y}_i - \bar{y})^2$$

- 만약 선형모델 가정이 맞다면 다음이 성립한다.

$$E[MSE] = \sigma^2$$
$$since\ SSE = \sum (y_i - \bar{y})^2 = \sum e_i^2$$

- 또한 귀무가설 H_0가 맞다면 다음이 성립한다.

$$E[MSR] = \sigma^2$$

- 설명변수들과 반응변수 사이에 상관관계가 없다는 귀무가설 H_0 하에서는 MSR이 MSR과 비슷하며, 즉 둘의 비가 1에 매우 가까울 것으로 기대되기 때문이다.
- 만약 대립가설 H_1이 참이라면 회귀모형은 유의미하고, SSR이 SST의 많은 부분을 설명하므로, MSR이 MSE보다 커진다. 즉 $E[MSR] > \sigma^2$이고 MSR과 MSE의 비는 1보다 크다.

* 분산분석에서의 F-test는 그룹 간 변동량(Between, MSTR)과 그룹 내 변동량(Within, MSE)을 비교함으로써 평균의 동일성을 검정하는 기법이다. 다중선형회귀에서의 F-test는 회귀에 의한 변동량(MSR)과 오차에 의한 변동량(MSE)를 비교함으로써 회귀모형의 적절성을 판단한다.
* F-통계량은 F분포를 따른다. 따라서 F-통계량이 (1에 비해) 얼마나 유의미하게 큰지 p값을 계산할 수 있으며, 이를 기준으로 기각 여부를 결정한다.
* 때로는 특정 q개 계수가 0인지를 검정하고자 할 때가 있다. 이 경우 마지막 계수 q개 계수를 제외한 모든 변수들을 사용하는 두 번째 모델이 사용된다. 다음의 F통계량으로 Reduced Model (p-q)에 비해 Full Model (p)이 얼마나 더 많은 변동량을 설명하는지에 대한 F-test를 실시할 수 있다.

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$
$$F^* = \frac{(SSE(R) - SSE(F))/q}{SSE(F)/(n-p-1)} \sim F(q, n-p-1)$$

* 각 설명변수에 대한 t-통계량과 p-값은 어떤 의미일까?
  - 각 설명변수가 다른 설명변수들을 조정한 후에 반응변수와 상관성이 있는지
  - Full Model에서 그 변수를 제외하고 나머지 변수들은 모두 포함하는 Reduced Model의 F-검정
  - 즉, q=1인 Reduced Model의 F-검정
  - 따라서 모델에 해당 변수를 추가하는 것에 대한 *부분적 효과*를 나타낸다.
* 각 변수에 대한 p-값이 있는데 왜 F-통계량을 살펴볼 필요가 있는가?
  - If any one of the p-values for the individual variables is very small, then at least one of the predictors is related to the response.
  - 이러한 결정방식은 결점이 있으며, 특히 설명변수의 개수 p가 클 경우에 그렇다.
  - "각 설명변수와 연관된 p-값들 중에서 약 5%는 우연히 0.05보다 작을 것이다." p-값이 유의수준 0.05보다 작다는 것은 오판됨을 의미한다. 다시 말해 5%의 설명변수들은 그 상관관계가 오판될 것이라는 뜻이다. 왜냐하면 개별 설명변수는 오판가능성 = Type 1 error를 유의수준 0.05 정도까지 갖고 있기 때문이다.
  - p=100이라면 100개의 설명변수들 중 5개, 실질적으로 적어도 1개는 오판된다.
  - 이처럼 실제로 상관관계가 없는데도 p-값 < 0.05이므로 상관관계가 있다고 결론을 내리는 경우가 생긴다.
* 그러나 F-통계량은 변수의 개수를 고려하므로 이러한 오판의 가능성을 줄여준다. 이는 총체적 검정이므로 p개의 설명변수들 각각의 t-통계량과 달리 단 하나이며, 오판 가능성은 5%이다.
  - p > n이면 이용할 관측치 수보다 추정할 계수 β_j가 더 많으므로 F-통계량을 이용할 수 없다.
  - p가 클 때에는 전진선택과 같은 방법을 사용할 수 있다.

둘: 중요 변수의 결정

- F-통계량과 관련된 p-값을 살펴보았더니, 적어도 하나의 설명변수는 반응변수가 상관성이 있었다.
- 그렇다면 그러한 변수가 어떤 것들인지 궁금할 것이다.
  - 각 설명변수의 t-통계량? 위에서 설명한 위험이 따르며, 특히 p가 큰 값인 경우 그렇다.
  - 대부분의 경우 설명변수들의 일부만이 반응변수와 상관관계가 있다.
  - 상관성이 있는 설명변수들만으로 모델 적합을 수행하기 위해, 어느 설명변수가 반응변수와 상관성이 있는지 결정하는 것을 **변수선택**이라고 한다.
- 이상적으로는 p개 변수들로 가능한 모든 모델을 시험하고, 그중 최고의 모델을 선택한다.
- 그러나 가능한 경우의 수는 2^p개이고, p의 값이 크다면 그 수는 비현실적으로 커진다.
- 따라서 2^p개보다 더 작은 수의 모델 집합을 고려 대상으로 하는 기법이 필요하다.
  - 3가지 고전적인 기법들

> 1. 전진선택

- 영모델로 시작해, 매 단계에서 가장 낮은 SSE가 생기는 변수를 추가한다.
- 항상 사용할 수 있지만, 초기에 포함한 변수들이 나중에는 유효하지 않을 수 있다.

> 2. 후진선택

- 모든 변수를 가지고 시작해, 매 단계에서 가장 p-값이 큰 변수를 제외한다.
- p>n이면 사용할 수 없다.

> 3. 혼합선택

- 전진선택과 후진선택을 결합한 것이다.
- 영모델로 시작해, 유의미한 변수를 추가하고, 어떤 하나의 p-값이 너무 커지면 제외한다.

셋: 모델 적합

- 단순선형회귀에서와 같은 방식으로 RSE와 R^2가 수치적 측도로 사용된다.

> R^2

- 단순회귀에서는 반응변수와 설명변수의 상관변수의 제곱이었다.
- 다중선형회귀에서는 반응변수와 적합된 선형모델 사이의 상관계수 제곱인 $Cor(Y, \hat{Y})$와 같다.
- 1에 가까우면 반응변수 내 분산의 많은 부분을 설명한다.
- 그런데 모델에 더 많은 변수가 추가되면 (훈련 데이터에 대한) R^2은 항상 증가할 것이다.
- 만약 설명변수를 추가했는데도 R^2가 약간만 증가한다는 것은 이것이 모델 적합을 특별히 개선시키지 못하며, 오히려 과적합 문제를 발생시킬 가능성이 높다는 것을 의미한다.

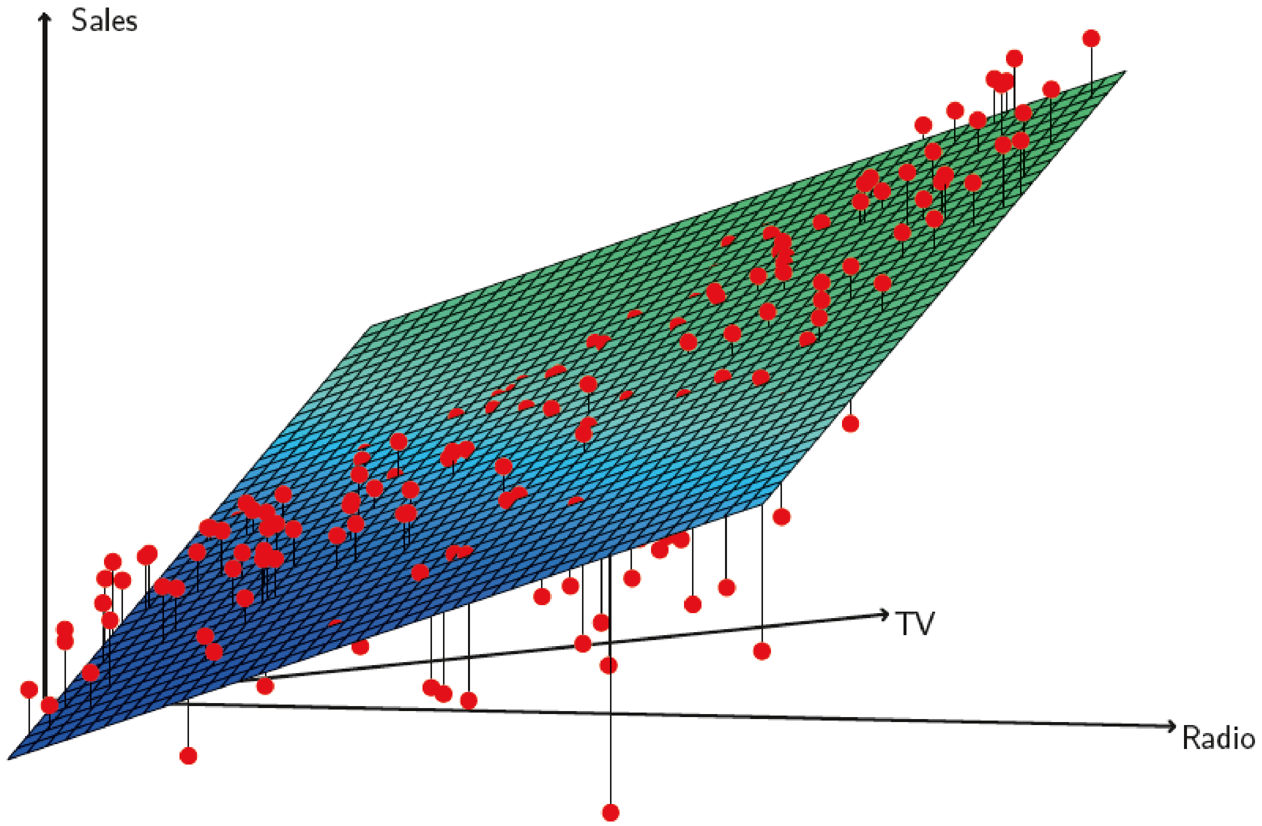
> RSE (잔차표준오차)

- How RSE can increase when newspaper is added, given that RSS must decrease?

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

- 더 많은 변수를 가진 모델은 RSS 감소량이 p 증가에 비해 상대적으로 작을 경우 RSE가 높아진다.
- 아래 그림은 수치적으로 표현되지 않는 모델의 문제점을 보여준다.
  - 선형모델은 광고예산이 TV나 라디오 어느 한 쪽에 편중되면 판매량을 과대추정한다.

- 선형모델은 예산이 분산될 때에는 판매량을 과소추정한다.
- 즉 비선형 상관관계가 뚜렷하며, 이는 광고매체 간에 *시너지 또는 상호작용 효과*가 있기 때문이다.
- 상호작용 항을 이용하여 시너지 효과를 수용하는 것은 3.3.2절에서 다룬다.



넷: 예측

- 설명변수 $X_1, X_2, ..., X_p$의 값에 기초하여 반응변수 Y를 예측한다.
  1. 최소제곱평면 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_p$은 실제 모회귀평면 $f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$에 대한 추정값이다. 계수추정의 부정확도는 축소가능 오차와 관련된다.
  2. *모델 편향(model bias)*이라는 잠재적인 축소가능 오차가 있다. 선형모델이라는 것 역시 현실에 대한 근사에 불과하다. 다른 모델이 최상의 모델일 수 있으나, 이는 무시하고 선형모델이 올바른 것으로 간주한다. 즉 우리는 "최상의 선형 근사"라는 한계 안에서 "선형 모델"을 추정하는 것이다.
  3. 정확한 모수값을 알더라도 모델의 랜덤오차 때문에 반응변수 값을 완벽하게 예측할 수 없다. 이를 축소불가능 오차라고 하였다.
- Y는 $\hat{Y}$와 얼마나 다를 것인가?
  - 예측구간(Prediction Interval)은 신뢰구간(Confidence Interval)보다 항상 더 넓다.
  - 예측구간은 $f(X)$에 대한 추정오차 = 축소가능 오차, 각 포인트가 모회귀평면과 얼마나 다른지에 대한 불확실성 = 축소불가능 오차 둘 다 포함하기 때문이다.
  - 신뢰구간은 mean response에 대한 불확실성을 수량화하는 데, 예측구간은 individual outcome에 대한 불확실성을 수량화하는 데 사용된다.
  - 신뢰구간은 축소가능 오차만 포함하고, 축소불가능 오차는 포함하지 않는다.

3.3 회귀모델에서 다른 고려할 사항

3.3.1 질적 설명변수

- 지금까지는 선형회귀모델의 모든 변수는 양적이라고 가정하였다.
- 하지만 실제로는 설명변수들이 질적인 경우도 많다.

  > 레벨 수가 2인 설명변수

  - 단순히 두 개의 값을 갖는 지시변수, 또는 *가변수*를 생성한다.

$$x_i = \begin{Bmatrix} 0\ (if\ female) \\ 1\ (if\ male) \end{Bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{Bmatrix} \beta_0 + \beta_1 + \epsilon_i\ (if\ female) \\ \beta_0 + \epsilon_i\ (if\ male) \end{Bmatrix}$$

  - 가변수를 어떻게 만드는지에 따라 계수들을 해석하는 방식이 달라진다.

$$x_i = \begin{Bmatrix} 1\ (if\ female) \\ -1\ (if\ male) \end{Bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{Bmatrix} \beta_0 + \beta_1 + \epsilon_i\ (if\ female) \\ \beta_0 - \beta_1 + \epsilon_i\ (if\ male) \end{Bmatrix}$$

  - 여성이 남성에 비해 신용카드 대금이 평균적으로 얼마나 높은지 vs. 여성은 (여성과 남성의) 평균에 비해 얼마나 높고, 남성은 평균에 비해 얼마나 낮은지

  > 레벨 수가 3 이상인 질적 설명변수

  - 레벨 수가 3인 경우 다음과 같이 2개의 가변수를 생성한다.

$$x_{i1} = \begin{Bmatrix} 0\ (if\ Asian) \\ 1\ (if\ not\ Asian) \end{Bmatrix}$$

$$x_{i2} = \begin{Bmatrix} 0\ (if\ White) \\ 1\ (if\ not\ White) \end{Bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \epsilon_i = \begin{Bmatrix} \beta_0 + \beta_1 + \epsilon_i\ (if\ Asian) \\ \beta_0 + \epsilon_i\ (if\ male) \\ \beta_0 + \beta_1 + \epsilon_i\ (if\ Asian) \end{Bmatrix}$$

  - 가변수의 개수는 항상 레벨 수보다 하나 작을 것이다.
  - 가변수가 없는 레벨은 *기준(baseline)*으로 알려져 있다.
  - 개별 계수에 의존하지 않고 F-검정을 사용하여 $H_0 : \beta_1 = \beta_2 = 0$을 검정할 수 있다.
- 이런 가변수 방식은 양적 설명변수와 질적 설명변수를 둘 다 포함하는 경우에 사용할 수 있다.

3.3.2 선형모델의 확장

- 선형회귀모델은 실제로 성립되지 않는 몇 가지 제한적인 가정을 사용한다.
- 가장 중요한 가정 중 두 가지는 설명변수와 반응변수 사이의 관계가 **가산적**이고 **선형적**이라는 것이다.
  - 가산성(additive): 설명변수 X_j의 변화가 반응변수 Y에 미치는 영향은 다른 설명변수 값에 독립적
  - 선형성: X_j의 한 유닛 변화로 인한 Y의 변화는 X_j의 값에 관계없이 상수
- 이러한 두 가정을 완화시키는, 선형모델을 확장시키는, 고전적 기법에 대해 알아볼 것이다.

> 가산성 가정의 제거

- 예를 들어, 선형모델은 TV 지출의 한 유닛 증가가 판매량에 미치는 평균 영향은 라디오 광고 지출액에 관계없이 β_1이라는 것을 의미한다. 그러나 이런 단순한 모델은 시너지 효과 또는 상호작용 효과를 고려하지 않는다.
- 상호작용 효과를 포함하도록 이 모델을 확장하는 한 가지 방법은 *상호작용 항*을 포함하는 것이다.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \\
&= \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\
&= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned}
$$

- 상호작용항인 TV x radio의 p-값은 매우 작고 R^2는 커졌으므로, 주효과만 포함하는 모델에 비해 훨씬 낫다.
- 만약 X_1과 X_2 사이의 상호작용이 유의미하다면 X_1과 X_2 각각의 주효과가 유의하지 않더라도 포함한다.
  - 상호작용항이 반응변수와 상관관계가 있으면 X_1 또는 X_2의 계수가 0인지는 관심이 없다.
- 상호작용의 개념은 질적 변수 또는 양적 변수와 질적 변수의 조합에도 적용된다.
  - 소득 변화가 신용카드 대금에 미치는 영향이 학생인지의 여부에 따라 다를 수 있게 한다.
  - 소득 증가에 따른 카드 대금 증가가 학생인 경우 학생이 아닌 사람보다 낮다.

> 비선형 상관관계

- 선형회귀모델은 반응변수와 설명변수의 상관관계가 선형적이라고 가정했으나, 실제로는 비선형적일 수 있다.
- 여기서는 비선형 상관관계를 수용하도록 *다항식회귀*를 사용하여 선형모델을 확장한다.
- 이차항 이상의 변환된 형태의 설명변수들을 모델에 포함하여 곡선 상관관계를 설명한다.
- *그러나 이것은 여전히 선형모델이다!*
  - 단순히 $X_1 = horsepower$과 $X_2 = horsepower^2$를 갖는 다중선형회귀모델이다.
  - 이차적합의 R^2와 이차항 추정계수의 p-값을 고려했을 때 상당히 유의하다.
  - 더 고차항을 포함할 경우 지나치게 구불구불해질 수 있다.
- 다항식회귀로 불리는 이유는 회귀모델에 설명변수들의 다항식 함수를 포함하기 때문이다.

3.3.3 잠재적 문제

1. 반응변수-설명변수 상관관계의 비선형성

- 실제 상관관계가 선형과 거리가 멀면 적합 및 예측 정확도는 신뢰하기 어렵다.
- 이상적이라면 *잔차 그래프*는 인지할 만한 패턴을 보이지 않을 것이다.
- 만약 잔차 그래프가 비선형 상관성이 있다는 것을 나타내면, 설명변수들을 비선형적으로 변환한다.

2. 오차항들의 상관성

- 선형회귀모델에서 중요한 가정은 오차항들이 서로 상관되어 있지 않다는 것이다.
- 만약 오차항들 사이에 상관성이 있으면 추정된 표준오차는 실제 표준오차를 과소추정할 것이다.
  - 축소불가능 오차인 오차항의 분산에, 오차항들 간 상호작용으로 인한 분산이 추가되기 때문이다.
  - 그 결과 실질적인 신뢰구간과 예측구간은 계산된 수치보다 더 좁을 것이다.
- 즉 오차항이 상관되어 있을 경우 모델에 대한 확신에 근거가 부족해진다.
  - 실수로 데이터 값이 2배가 되었다면, 예상치 못하게 오차항들 사이에 상관성이 생긴 것이고, 그에 따라 표준오차가 표본크기가 2n인 것처럼 계산되어 신뢰구간도 좁아진다.
  - 오차항들 사이의 상관관계는 시계열 데이터에서 자주 발생한다. 이웃하는 시점에 얻어진 관측치들은 그 오차가 양의 상관성을 가질 것이기 때문이다.

3. 오차항의 상수가 아닌 분산

- 가정과 달리 오차항들의 분산은 상수가 아닐 것이다.
- 오차항의 비상수 분산 또는 이분산성(heteroscedasticity)은 잔차 그래프에 깔때기 형태로 알 수 있다.
- 예를 들어, 오차항들의 분산은 반응 변수의 값에 따라 증가할 수 있다.
- 이런 문제는 오목함수를 사용하여 반응변수 Y를 변환하는 것이 한 가지 해결책이다.
- *가중최소제곱(weighted least squares)*: 반응변수의 분산이 일정한 패턴을 보일 때
    - Let i^{th} observation (y_i) be an average of n_i raw observations (x_i).
    - If each x_i is uncorrelated with variance σ^2, then y_i has variance σ^2/n.
    - In this case, we can fit our model by giving weight proportional to the inverse variance.
    - $i.e., w_i = n_i$

4. 이상치 (outlier)

- 이상치는 y_i가 모델이 예측한 값과 크게 다른 점이다.
- 이상치는 최소제곱적합에 큰 영향을 미치지는 않아도, SSE나 R^2에 영향을 미친다.
    - 이상치는 레버리지가 높은 경우와 달리 설명변수 값이 특이한 것이 아니다.
    - 따라서 최소제곱선에 거의 영향을 주지 않는다.
- 잔차 그래프, 또는 이상치의 판단 기준(절댓값이 3)을 제공하는 스튜던트화 잔차를 그릴 수 있다.

5. 레버리지가 높은 (영향력이 큰) 관측치

- 이상치는 주어진 설명변수의 값 x_i에 대해 반응변수의 값 y_i가 보통 수준과 다르다.
- 높은 레버리지를 갖는 관측치는 x_i값이 보통 수준과 다르다.
    - 레버리지(leverage): 영향력, 지렛대
- 이상치와 달리 높은 레버리지 관측치는 추정회귀선에 상당한 영향을 주기 때문에 식별이 중요하다.
- 그러나 단순선형회귀에서는 이를 찾기 쉽지만, 다중회귀에서는 여러 변수와 그것들의 차원을 고려해야 하므로 식별하기 어렵다.
- 따라서 관측치의 레버리지를 수량화하기 위해 *레버리지 통계량*을 계산한다.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- 위는 단순선형회귀의 경우이며, 다중회귀로 확장하면 평균 레버리지는 항상 $(p+1)/n$이다.
- 따라서 $(p+1)/n$보다 큰 레버리지 통계량을 갖는 점은 높은 레버리지를 갖는다.

6. 공선성 (Collinearity)

- 공선성은 두 개 또는 그 이상의 설명변수들이 서로 밀접하게 상관되어 있는 경우를 말한다.
- 공선성이 있을 경우 반응변수에 대한 공선형 변수들의 개별 효과를 분리하기 어려울 수 있다.
    - 회귀계수 추정치가 좁고 긴 계곡을 따라 어디로든 움직일 수 있다, 즉 정확성을 낮춘다.
    - 정확성이 낮아지므로 $\hat{\beta}_j$에 대한 표준오차는 감소, t-통계량은 증가, p-값은 증가한다.
    - 가설검정의 능력 - 0이 아닌 계수를 정확하게 검출할 확률 - 이 공선성에 의해 줄어든다.
- 공선성을 검출하는 간단한 방법은 설명변수들의 상관행렬을 살펴보는 것이다.
    - 상관행렬의 절댓값이 큰 원소는 상관성이 높은 변수들의 쌍을 나타낸다.
- 유감스럽게도 모든 공선성 문제가 상관행렬에 의해 발견되는 것은 아니다.

- *다중공선성(multicollinearity)*: 변수 쌍은 특별히 상관성이 없더라도 세 개 또는 그 이상의 변수들 사이에서는 공선성이 존재할 수도 있다.
- 상관행렬을 검사하는 대신 다중공선성을 판단하는 더 좋은 방법은 *분산팽창인수 (VIF, variance inflation factor)* 를 계산하는 것이다.

- VIF는 β_j에 대해 full model 적합의 분산을 reduced model 적합의 분산으로 나눈 것이다.
  - 아래의 공식으로 VIF를 계산할 수 있다.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

  - $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.
  - $R^2_{X_j|X_{-j}}$ 이 1에 가까우면 공선성이 존재하고, 따라서 VIF 값이 클 것이다.
  - VIF는 최소 1이며, 이는 공선성이 전혀 없음을 나타낸다.
  - VIF가 5 또는 10을 초과하면 문제의 소지가 있는 공선성을 나타낸다.
- 공선성을 해결하기 위해서는, (1) 한 변수를 제외하거나 (2) 새로운 설명변수로 결합하는 것이다.

3.4 마케팅 플랜

1. 광고예산과 판매 사이에 상관관계가 있는가?T
   - TV, radio, newspaper에 따른 sales의 다중회귀모델
2. 광고예산과 판매 사이에 얼마나 강한 상관관계가 있는가?
   - RSE, R^2 통계량을 살펴본다.
3. 어느 매체가 판매에 기여하는가?
   - 각 설명변수의 t-통계량과 연관된 p-값을 조사한다.
4. 판매에 대한 각 매체의 효과는 얼마나 되는가?
   - β_j의 표준오차를 이용해 β_j의 신뢰구간을 구한다.
   - VIF를 구해 공선성의 증거가 없음을 확인한다.
   - 판매량에 대한 각 매체의 개별 상관성을 평가하기 위해 세 개의 다른 단순선형회귀를 실시한다.
5. 미래의 판매량에 대해 얼마나 정확하게 예측할 수 있는가?
   - 개별 반응변수 값을 예측한다면 예측구간, 평균 반응변수 값을 예측한다면 신뢰구간을 사용한다.
6. 상관관계가 선형적인가?
   - 잔차 그래프에는 패턴이 없어야 한다.
   - 비선형 상관관계가 발견된다면 이를 수용하기 위해 설명변수들을 변환할 수 있다.
7. 광고 매체 사이에 시너지가 있는가?
   - 비가산적 상관관계를 수용하기 위해 상호작용 항을 포함한다.

3.5 선형회귀와 K-최근접이웃의 비교

- 선형회귀는 $f(X)$를 선형함수 형태라고 가정하기 때문에 모수적 기법이다.
  - 추정해야 할 계수의 수가 적기에 적합하기 쉽다.
  - 계수들에 대한 해석이 간단하고 통계적 유의성을 쉽게 검정할 수 있다.

- 실제 함수형태가 선형적이지 않고, 목적이 예측 정확도라면 모수적 방법은 적합하지 않다.

- 비모수적 방법은 $f(X)$에 대해 모수적 형태를 가정하지 않아 유연한 회귀 수행 기법이다.
  - 먼저 x_0에 가장 가까운 K개의 훈련 관측치 N_0를 식별한다.
  - 그 다음에 N_0 내의 모든 훈련 관측치들에 대한 반응변수 값들의 평균을 사용하여 $f(x_0)$를 추정한다.

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathrm{N}_0} y_i$$

  - K가 작으면 계단함수의 형태, K가 커질수록 적합은 더 평활해지고 변동(분산)은 줄어든다.
  - 그러나 평활화는 $f(X)$ 구조의 일부를 감춤으로써 편향을 초래할 수 있다.
  - 최적의 K값은 편향-분산 절충에 따라 다를 것이다.
- 모수적 방식은 선택된 모수 형태가 f의 실제 형태에 가까운 경우 비모수적 방식보다 낫다.
  - 비선형성의 정도가 증가하면 KNN이, 특히 K가 큰 경우, 선형회귀보다 낫다.
  - 현실에서는 실제 상관관계를 모르니 KNN을 사용해야 할까?
  - 상관관계가 비선형적인 경우 여전히 선형회귀가 나을 수 있고, 차원이 높으면 더욱 그렇다.
- 설명변수당 관측치의 수가 작으면 모수적 방법이 더 낫다.
  - 아래 그림에서 p=3을 기준으로 KNN이 선형회귀에 비해 검정 MSE가 확연히 커진다.
  - 차원의 저주: p가 너무 크면 가까운 이웃이 없으므로, 차원이 증가함에 따라 KNN은 성능이 나빠진다.
  - 차원이 낮은 경우에도 해석력 관점에서 선형회귀를 KNN보다 선호할 수 있다.