회 귀 분 석 I

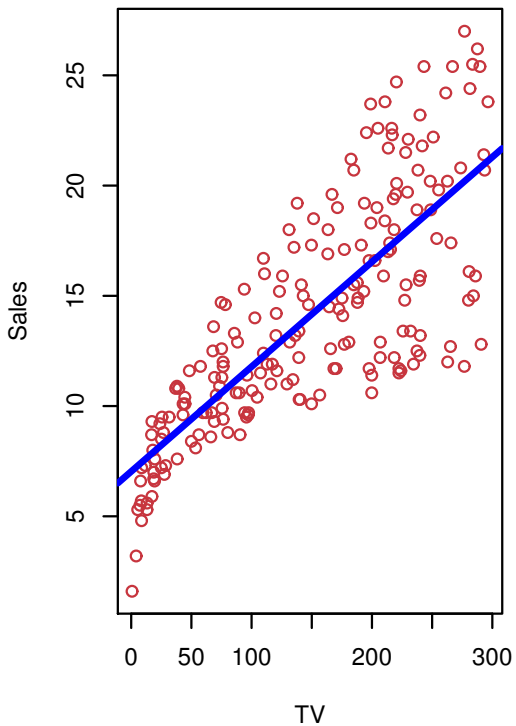
서 울 대 학 교
통 계 학 과

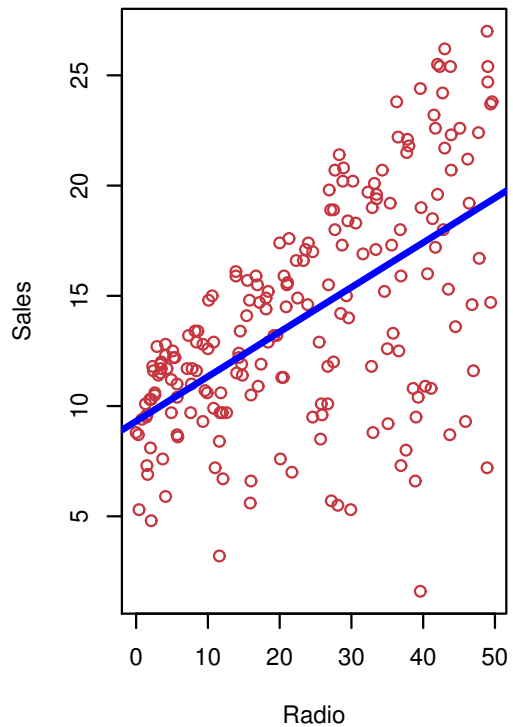
2018년 8월

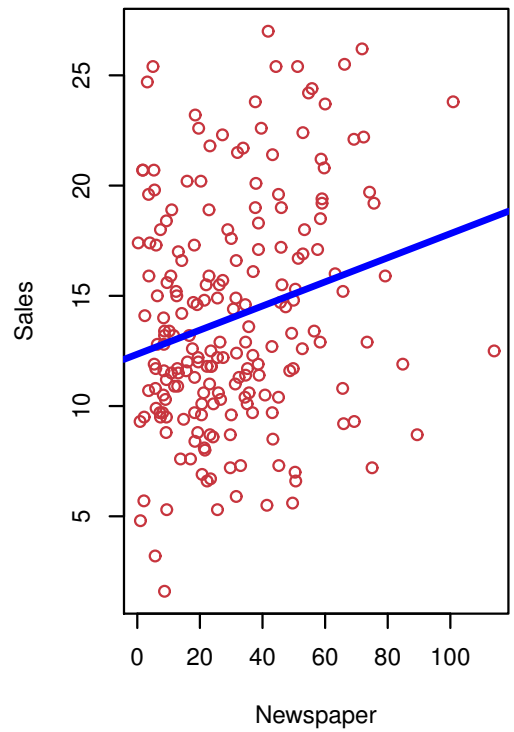
이 장에서 할 얘기들

1. 일반적 회귀모형의 소개 및 목적
2. 단순선형회귀
   - 2.1 회귀계수의 추정
   - 2.2 추정치의 성질 : 불편향성, 표준오차
   - 2.3 모형 적합성 척도 : RSE, R^2, F-통계량
3. 중회귀모형
   - 3.1 회귀계수의 추정
   - 3.2 잠복변수(lurking variable)
   - 3.3 모형 적합성 척도 : R^2, F-통계량
4. 가변수
5. 가정의 완화 : 가법성 가정, 선형성 가정
6. 모형진단

광고 (Advertising) 자료







노트.

1. 200의 중소도시(마켓)에 관해 매출(단위 천개), TV, radio, newspaper 광고 지출액(단위 천불)을 포함한 자료이다.
2. 그림은 매출과 광고지출액 간에 관계를 나타낸다.
3. 매출액을 크게 하고 싶지만, 우리가 직접 매출액을 조정할 수는 없다. 하지만, 티비, 라디오, 신문광고비용은 우리가 조정할 수 있다. 티비, 라디오, 신문광고비용이 매출에 어떤 영향을 미치는지 알면, 매출액을 간접 조정할 수 있다.

노트. 광고자료를 보고 할 질문들

질문은 크게 예측에 관한 질문과 관계에 관한 질문으로 나뉠 수 있다.

1. 예측에 관한 질문
   - 1.1 각 미디어의 매출에 대한 효과를 얼마나 정확히 추정할 수 있나? TV 광고에 한달에 만불을 지출하면 매출이 얼마나 오르나?
   - 1.2 미래의 매출액을 얼마나 정확히 예측할 수 있나? 주어진 TV, 라디오, 신문 광고비용을 지출하면 매출이 얼마가 되나?
2. 관계에 관한 질문
   - 2.1 광고예산과 매출 사이에 관계가 있나? 없다면 광고를 하지 말자고 주장할 수 있다.
   - 2.2 관계가 있다면 그 관계가 얼마나 강한가? 강한 관계가 있다면 사용된 광고비로 매출을 정확히 예측할 수 있고, 약한 관계라면 무작위추측(random guess)보다는 조금 더 나은 예측을 할 수 있을 것이다.
   - 2.3 어떤 미디어가 매출에 영향을 미치나? 이 질문에 답하기 위해서는 각 미디어의 효과를 분리할 수 있어야 한다.
   - 2.4 변수간 관계가 선형인가? 선형이면 선형회귀를 이용하고, 아니면 다른 방법을 이용해야 한다.
   - 2.5 광고매체사이에 시너지 효과가 있는가, 즉 교호작용(interaction effect)가 있는가?

회귀모형

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

Y : 반응변수, 종속변수

$X = (X_1, \ldots, X_p)$: 예측변수, 독립변수, 설명변수

ϵ : 평균이 0인 랜덤 오차항

회귀분석의 목적

예측(prediction)

$\hat{Y} = \hat{f}(X)$로 Y 값을 예측

추론(inference)

X와 Y의 관계를 이해

노트.

회귀분석의 목적은 크게 두 가지로 나눌수 있다. 두 가지 목적이 완전히 분리되지는 않는다.

예측(prediction)

$\hat{Y} = \hat{f}(X)$로 Y값을 예측. 내년에 TV 광고비를 3만불, 라디오 광고비 2만불을 쓸 예정인데, 매출이 얼마가 될까?

추론(inference)

X와 Y의 관계를 이해. e.g. 매출에 가장 큰 영향을 미치는 미디어는 무엇인가? TV 광고를 x 만큼 증가시키면 매출은 얼마나 증가하나?

단순선형회귀모형(simple linear regression model)

모형

$$Y = \beta_0 + \beta_1 X + \epsilon$$

β_0, β_1 : 회귀계수 ϵ : 오차항

예

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}.$$

예측

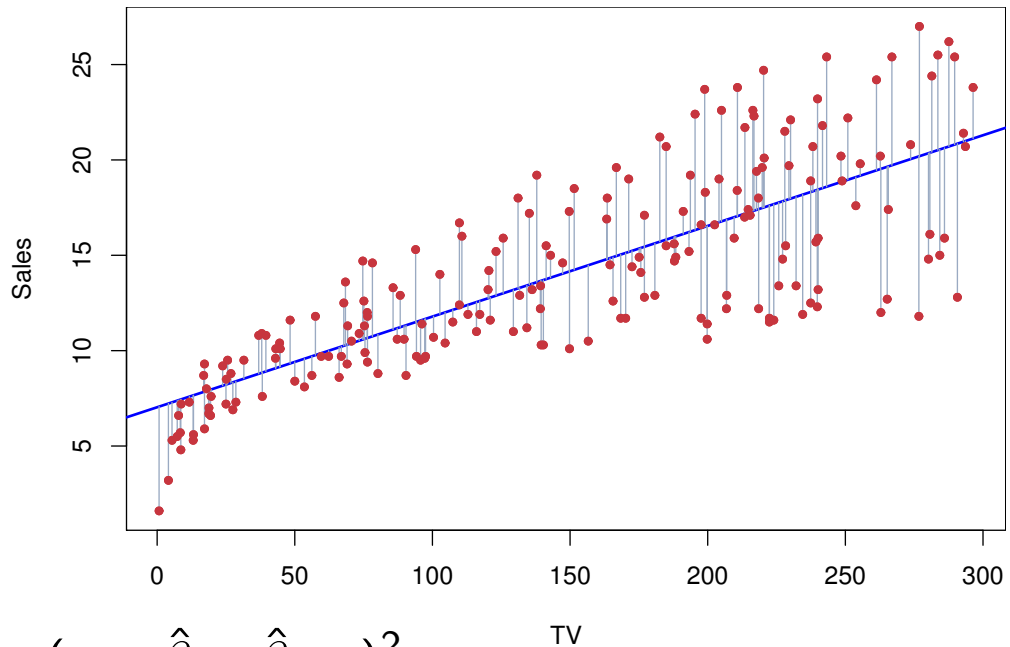
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

노트.

1. 회귀모형 중 가장 간단한 모형이다.
2. 직선의 식에서 β_0는 y 절편, β_1은 기울기 이다.
3. ϵ은 $\beta_0 + \beta_1 X$로 설명하지 못하는 모든 것이다.
4. 자료를 이용해 회귀계수의 추정지 $\hat{\beta}_0, \hat{\beta}_1$을 구하면 다음의 식을 이용해 예측을 할 수 있다.

회귀계수의 추정 : 최소제곱법

최소제곱법(least squares method) : 잔차제곱합을 최소화하는 $\hat{\beta}_0, \hat{\beta}_1$을 구하는 방법



잔차제곱합 (RSS, residual sum of squares)

$$(y_1-\hat{\beta}_0-\hat{\beta}_1x_1)^2+(y_2-\hat{\beta}_0-\hat{\beta}_1x_2)^2+\ldots+(y_n-\hat{\beta}_0-\hat{\beta}_1x_n)^2$$

잔차(residual)

$$e_i = y_i-\hat{y}_i = y_i-(\hat{\beta}_0+\hat{\beta}_1x_i),\ i = 1,2,\ldots,n.$$

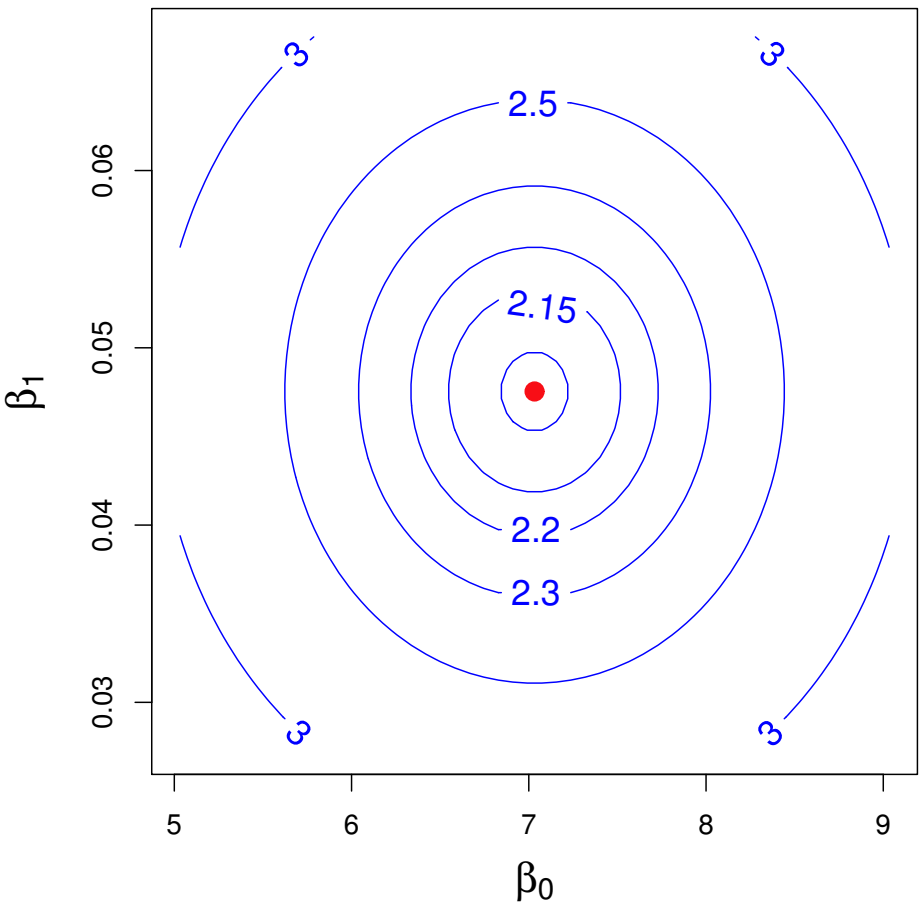
노트.

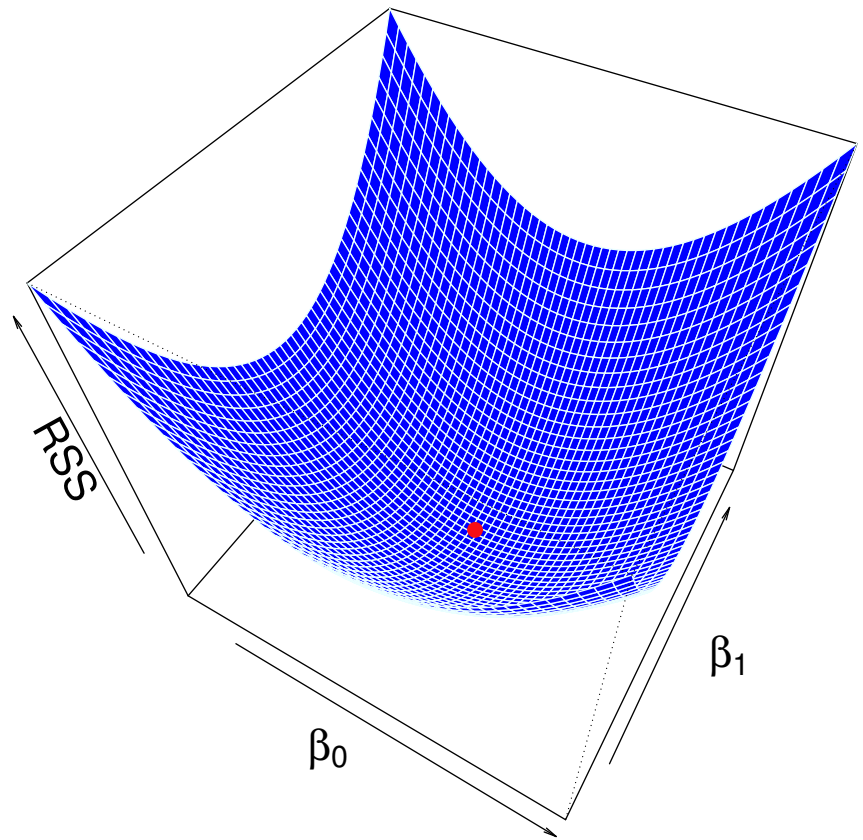
1. **목적.** 자료 $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$이 주어져 있을 때, $\hat{\beta}_0, \hat{\beta}_1$을 구하려고 한다. (β_0, β_1) 한 쌍은 한 개의 직선을 표현한다. (β_0, β_1)값을 추정하는 것은 자료를 가장 잘 요약하는 직선을 찾아내는 것이다.
2. **최소제곱법(least squares method)이란.**
   잔차제곱합(RSS, residual sum of squares)

   $$(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

   을 최소화하는 $\hat{\beta}_0, \hat{\beta}_1$을 구하는 것이다.
   최소제곱법은 회귀계수를 추정할 때 가장 많이 사용하는 방법이다.
3. 그림은 잔차를 그림으로 보여준다.
4. 자료에서 벗어난 직선은 잔차가 커지게 되어 있다.

최소제곱법





노트.

1. 그림은 잔차제곱합의 등고선 그림이다. 잔차제곱합과 추정치의 관계를 보여준다.
2. 잔차제곱합은 β_0, β_1에 관한 2차식이다.
3. 잔차제곱합의 표면그림이다. 잔차제곱합 표면에 추정치가 위치한 것을 보여준다.
4. 그림에서 점 (β_0, β_1)은 한 개의 직선을 나타낸다. 직선의 기울기와 절편값을 조금씩 변하게 하면 잔차제곱합이 어떻게 바뀌는지 보여주는 그림이다.

최소제곱합

추정치의 식

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

$$\bar{x}=\frac{1}{n}\sum_{i1=}^{n}x_i, \quad \bar{y}=\frac{1}{n}\sum_{i1=}^{n}y_i$$

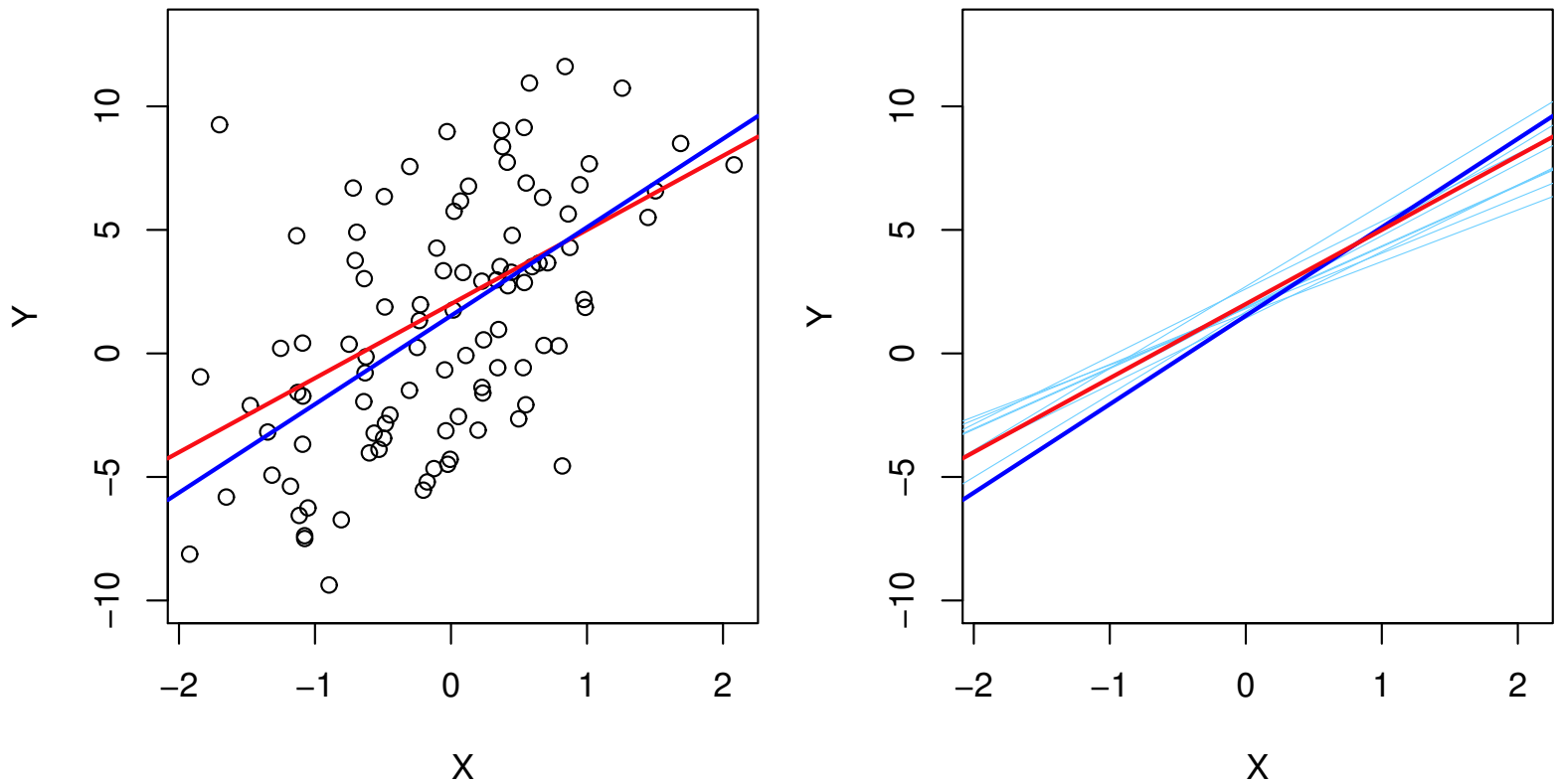
광고 자료 예

$$\text{sales} = 7.03+0.0475\times \text{TV}$$

노트.

1. 이렇게 구해진 추정치는 다음과 같이 주어진다. 잔차제곱합이 β_0, β_1에 대해 이차식이기 때문에 이를 최소로 하는 값을 식으로 구할 수 있다.
2. TV 광고에 1000불이 더 사용되면 매출은 47.5개 늘어난다.

추정치의 변동



노트.

1. 그림의 왼쪽 그림은 100개의 관측치를 생성해서 추정한 회귀직선을 나타낸다. 붉은 색은 진짜 회귀직선을 검은색은 추정된 회귀직선을 나타낸다. 오른쪽 그림은 100개의 관측치를 가진 자료를 여러번 생성해서 각 자료를 이용해 추정한 회귀직선들을 그렸다. 붉은색 진짜 회귀직선 주변에 추정된 직선들이 모여있다.
2. 실제 자료분석에서 우리는 진짜 회귀직선도 모르고 여러 개의 추정된 회귀직선도 알지 못한다.
3. 우리가 추정한 추정치가 참값을 중심으로 모여있는지, 엉뚱한 값을 중심으로 모여있는지, 모여있다면, 가깝게 모여있는지, 멀리 모여있는지 알고자 한다. 이를 알아야 추정치에 대해 자신감을 가질 수 있다. 이를 다루는 개념이 불편성과 표준오차이다. 다음 쪽에서 다룬다.

불편향성과 표준오차

불편성

$$\begin{aligned} \mathbb{E}(\hat{\beta}_0) &= \beta_0 \\ \mathbb{E}(\hat{\beta}_1) &= \beta_1 \end{aligned}$$

표준오차

$$\begin{aligned} SE(\hat{\beta}_0)^2 &= \mathbb{V}ar(\hat{\beta}_0) \\ &= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \end{aligned}$$

$$\begin{aligned} SE(\hat{\beta}_1)^2 &= \mathbb{V}ar(\hat{\beta}_1) \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

σ의 추정

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-2}}$$

노트.

1. 오차항 ϵ들간에 자기상관성이 없고 평균이 0 분산이 σ^2라는 가정하에서 위 식이 얻어졌다.
2. 불편성 혹은 불편향성은 추정치들이 참값을 중심으로 모여있는지 나타내고 표준오차는 추정치가 추정치들의 중심에서 얼마나 떨어져 있는지 나타낸다.
3. 최소제곱합으로 구해진 β_0와 β_1의 추정량은 불편향성을 가진다. 즉 참값 중심으로 모여있다.

신뢰구간과 가설검정

95% 신뢰구간

$$\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1)$$
$$\hat{\beta}_0 \pm 1.96 SE(\hat{\beta}_0)$$

p -값(p-value)이란

H_0가 참일 때 현재의 관측치보다 더 H_1에 가까운 혹은 H_1을 지지할 자료를 관측할 확률이다.

$$\text{p 값} = \mathbb{P}_{H_0}(|T| \geq |t|)$$

여기서 t는 관측된 t 값이다.

가설

H_0 : X와 Y 사이에 관계가 없다. vs H_1 : X와 Y 사이에 관계가 있다.

$$\Longrightarrow H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0.$$

유의수준 (significance level) $100\alpha\%$ 일 때 기각역

$$|t| = \left|\frac{\hat{\beta} - 0}{SE(\hat{\beta}_1)}\right| > t_{\alpha/2}(n-2)$$

α는 0에서 1사이의 값인데, 보통 0.01, 0.05, 0.1 중 하나를 쓴다.

노트.

1. p 값이 매우 작다는 것은 H_0가 참일 때 관측되기 어려운 값을 관측했다는 것이다. 따라서 이는 H_1을 지지하는 증거이다.
2. 1.96은 n이 클 때의 근사값이다. $t_{0.025}(n-2)$가 보다 정확한 값이다.
3. 표준오차값을 알면 신뢰구간을 구하고, 가설검정을 할 수 있다.

$sales = \beta_0 + \beta_1 \times TV$ 를 적합했을 때의 결과

```
> lm.fit <- lm(sales ~ tv, data=advertising)
> summary(lm.fit)

Call:
lm(formula = sales ~ tv, data = advertising)
Residuals:
    Min      1Q  Median      3Q     Max
-8.3860 -1.9545 -0.1913  2.0671  7.2124

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.032594   0.457843   15.36   <2e-16 ***
tv          0.047537   0.002691   17.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 3.259 on 198 degrees of freedom
Multiple R-squared:  0.6119,Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
> confint(lm.fit, level=0.95)
                 2.5 \%     97.5 \%
(Intercept) 6.12971927 7.93546783
tv          0.04223072 0.05284256
```

노트.

1. confint는 추정된 회귀계수들의 신뢰구간을 준다.
   위에서는 95% 신뢰구간이 주어져 있다.

$sales = \beta_0 + \beta_1 \times TV$ 를 적합했을 때의 결과

	회귀계수	표준오차	t-통계량	p값
intercept	7.0325	0.4578	15.36	≤ 0.0001
TV	0.0475	0.0027	17.67	≤ 0.0001

노트.

1. 앞에서 R 결과로 주어진 결과를 정리한 것이다.
2. 두 변수 모두 p 값이 매우 작아 유의하다.

모형의 적합성

모형 적합성의 정도를 나타내는 척도

1. 잔차표준오차 : 3.26
2. R^2 : 0.612
3. F 값 : 312.1

잔차표준오차
residual standard error, RSE
root mse

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

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

노트.

1. 오른쪽의 값들은 광고 자료의 경우이다. 아래에서 잔차표준오차와 R^2에 대해서만 설명한다.F 값에 대해서는 나중에 설명한다.
2. **잔차표준 오차의 의미**
   - 2.1 광고자료의 경우 3.26이다. 오차항 σ의 추정치가 3.26이란 얘기다. β_0와 β_1을 정확히 알아도 우리 예측치의 표준오차는 3.26이라는 얘기다. 즉 95% 예측구간을 생각하면 약 $\pm 2 * 3.26 = \pm 6.52$의 오차를 갖는다.
   - 2.2 매출이 50,000개가 될 것이라고 예측했다하자. 예측하는데 β_0, β_1을 참값을 썼다할지라도 실제 매출은 $50,000 \pm 6,520$개가 될 것이다.
3. RSE는 σ의 추정량이다. root mse라고 쓰기도 한다.

R^2

정의

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 값은 X와 Y의 선형관계의 정도를 측정한다.

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$r = corr(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i -}$$

$$R^2 = r^2$$

R^2는 y의 전체 변동량 중 x를 이용해 선형모형으로 설명하는 비율이다.

노트.

1. R^2는 y의 전체 변동량 중 x를 이용해 선형모형으로 설명하는 비율이다. $0---1$ 사이의 값을 갖는다. %를 이용해 표현하기도 한다.
2. TSS는 y_i 값들이 y의 평균을 기준으로 얼마나 변동이 있는가를 나타내고, RSS는 y_i들이 $\hat{y}_i$로부터 얼마나 떨어져 있는가를 나타낸다. TSS-RSS는 $\hat{y}_i$ 값들이 y의 평균을 기준으로 얼마나 떨어져 있는가를 나타내고, 이는 우리가 모형으로 설명하는 y의 변동이다.
3. 광고자료의 경우 약 60%의 y의 변동이 x에 의해 설명된다.

1. **R^2 값 해석의 주의** 보통 R^2 값이 1에 가까우면 좋은 모형이고 0에 가까우면 나쁜 모형이라고 해석한다. 물리학 자료와 같이 σ가 매우 작은 경우는 이와 같은 해석이 무리가 없다. 이 경우 R^2 값이 작다는 것은 틀린 모형을 적합했다는 것이다. 그러나 사회학, 심리학과 같은 분야의 자료는 σ가 매우 커서 참모형을 적합했을 때도 R^2 값이 0에 가까운 값이 나올 수 있다.

중회귀모형

모형

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

X_j : j번째 예측변수
β_j : X_j의 회귀계수
ϵ : 오차

예

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

예측식

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

노트.

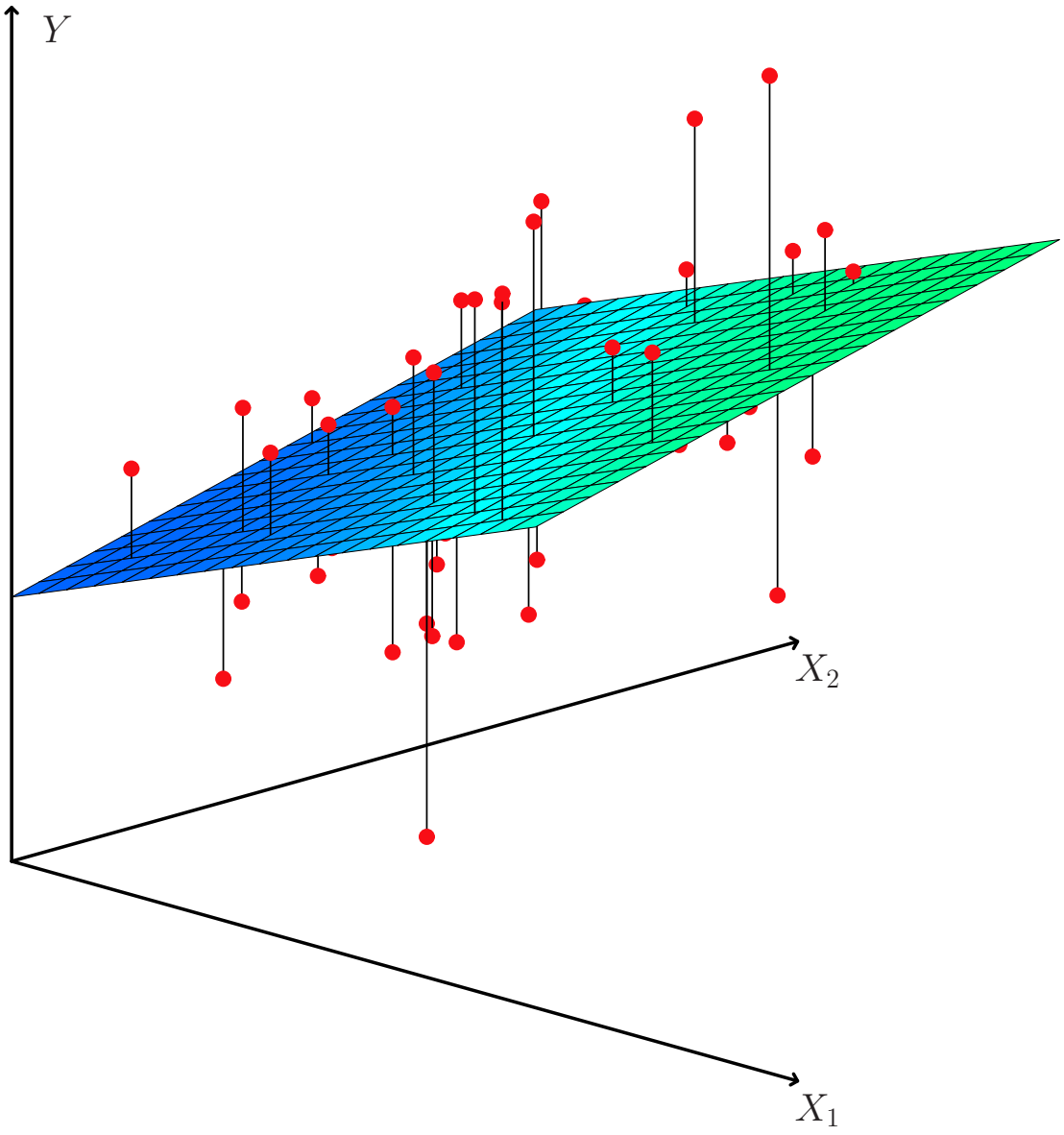
1. **중회귀모형** 예측변수가 여러 개 있을 대 Y의 예측식이 예측변수의 선형함수로 표현되는 모형.
2. β_j 다른 변수들의 값을 고정시키고 X_j가 1 만큼 변할 때 Y의 변화량. 좀 더 정확히는 $\mathbb{E}Y$의 변화량.

회귀계수의 추정

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

를 최소화하는 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$를 추정량으로 한다.



1. 그림은 잔차와 회귀평면의 그림을 보여준다.

$sales = \beta_0 + \beta_1 \times tv + \beta_2 \times radio + \beta_3 \times newpaper$ 적합 결과

```
> lm.fit <- lm(sales ~ tv + radio + newpaper, data=advertising)
> summary(lm.fit)

Call:
lm(formula = sales ~ tv + radio + newpaper, data = advertising)
Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
tv           0.045765   0.001395  32.809   <2e-16 ***
radio        0.188530   0.008611  21.893   <2e-16 ***
newpaper    -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

p 값이 큰 회귀계수의 해석

	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

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

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

질문

왼쪽표를 보면 newspaper의 회귀계수는 매우 작고 p 값은 매우 크다. 한 개의 변수만 모형에 포함시킨 오른쪽 아래의 표를 보면 newspaper의 회귀계수는 보다 크고 p 값은 유의하게 나온다. 어떻게 해석해야 할까?

노트.

1. **답변** TV, 라디오, 신문 3 개의 설명변수가 들어간 모형에서 TV와 라디오를 고정시키고 신문의 광고를 1000불 늘렸을 때 매출의 변화량은 -0.001이고 이 값이 0이 아니라는 강한 증거는 없다. 한편, 설명변수가 신문 하나만 들어간 모형의 결과에 대한 해석은 TV와 라디오 광고가 마음대로 변화할 때 신문 광고를 1000불 증가 시키면 매출이 0.055개 늘어난다는 뜻이다. 한 편, 다음 쪽에 있는 테이블의 상관계수를 살펴보면 신문과 라디오의 광고비용의 상관계수가 0.35로 라디오의 광고비용이 클 때 신문의 광고비용도 커지는 경향이 있다. 따라서 신문광고가 증가할 때 매출이 증가하는 현상은 실제로 신문광고가 증가해서 매출이 증가한 것이 아니라, 라디오 광고가 증가해서 매출이 증가한 현상을 나타낼 수 있다. 신문광고만 넣었을 때, 매출이 증가한 양은 신문 때문에만 변한 것인지, 라디오 때문에 변한 것인지 알 수 없다.

변수들 사이의 상관계수

```
> cor(adv)
                    tv       radio  newspaper     sales
tv          1.00000000 0.05480866 0.05664787 0.7822244
radio       0.05480866 1.00000000 0.35410375 0.5762226
newspaper   0.05664787 0.35410375 1.00000000 0.2282990
sales       0.78222442 0.57622257 0.22829903 1.0000000
```

노트. 변수들 사이의 상관계수

	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

노트.

1. **상어 공격 피해수 과 아이스크림 매출** 아이스크림 매출이 커지면 상어 공격 피해수가 커지는 현상이 있다. 이 자료를 이용해 상어 공격 피해수를 줄이기 위해서 해변에서 아이스크림 판매를 금지해야 한다는 주장이 있다면 말이 되는가? 보다 중요한 변수는 온도이다. 온도가 높아지면 해변의 방문객 수가 커지고 이에 따라 이이스크림 매출도 커지고 상어 공격 피해수도 늘어나는 것이다.

반응변수와 예측변수들 사이에 관계가 있는가?

가설

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ vs

H_1 : 최소한 한 개의 β_j는 0이 아니다.

F-통계량

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

노트.

1. **중회귀분석을 할 때 변수 선택에 관한 질문들**
   - 1.1 주어진 예측변수들 중 한 개의 변수라도 반응변수를 예측하는데 도움이 되나?
   - 1.2 $X_1, \ldots, X_p$ 중 어떤 예측변수들이 반응변수를 설명하는데 도움이 되나?
   - 1.3 모형이 자료를 얼마나 잘 적합하나?
   - 1.4 예측변수의 값들이 주어져 있을 대 반응변수의 값들을 어떻게 예측해야 하나? 또 이 예측값은 얼마나 잘 예측하나?

노트.

1. **F-통계량의 해석** 위의 선형모형이 사실이라면 $\mathbb{E}RSS/(n-p-1) = \sigma$ 이고, H_0 가 사실이라면

$$\mathbb{E}(TSS - RSS)/p = \sigma^2$$

이고 H_0 가 사실이 아니라면

$$\mathbb{E}(TSS - RSS)/p > \sigma^2$$

이다. 따라서 H_0 가 사실이라면 $F \approx 1$ 이고 귀무가설이 사실이 아니라면 $F >> 1$ 이 된다.

q개의 회귀계수가 0이라는 검정

가설

$H_0 : \beta_{p-q+1} = \ldots = \beta_p = 0$

vs $H_1 : \beta_{p-q+1}, \ldots, \beta_p$ 중 하나는 0이 아니다.

F-통계량

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

여기서 RSS_0는 q 개의 변수를 제외하고 적합한 회귀모형의 잔차제곱합이다.

노트.

1. **F 통계량과 t 통계량** $q = 1$일 때의 F 통계량은 이 예측변수의 t 통계량을 제곱한 값과 같다. 결국 두 개의 검정은 동일한 검정이다.
2. q개의 회귀계수가 0이라는 가설에 대한 F검정 코드를 넣는 것을 고려해보았는데, 이 검정이 많이 쓰이는 검정이 아니어서 넣지 않기로 했다.

변수선택: 어떤 변수가 중요한가?

모든 모형의 비교 방법

1. Mallow's Cp, AIC, BIC, adjusted R^2 등이 있다.
2. $p = 30$만 되어도 이 방법을 쓰기가 어렵다. $2^{30} = 1,073,741,824$.

근사적 방법

1. 전진선택법(forward selection)
2. 후진선택법(backward selection)
3. 단계적선택법(stepwise or mixed selection)

모형 적합성의 정도 : R^2

1. R^2는 Y의 변동 중 X로 설명되는 변동의 비율
2. $R^2 = Corr(Y, \hat{Y})^2$

신용카드자료

balance : 신용카드 대출의 크기

age : 나이(햇수)

cards : 신용카드의 개수

education : 교육받은 햇수

income : 수입(천불)

limit : 신용카드 한도

rating : 신용평가

gender : 성별

student : 학생신분 유무

status : 결혼 상태

ethnicity : 백인, 흑인, 동양인

노트.

1. 대출의 크기(balance)와 다른 변수들의 관계가 관심이 있다.

가변수(dummy variable) : 두 개의 값을 갖는 변수

가변수

$$x_i = \begin{cases} 1, & \text{i번째 사람이 여성} \\ 0, & \text{i번째 사람이 남성} \end{cases}$$

회귀모형에서 가변수의 해석

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{i번째 사람이 여성} \\ \beta_0 + \epsilon_i, & \text{i번째 사람이 남성} \end{cases} \end{aligned}$$

$balance = \beta_0 + \beta_1 \times genderFemale$

```
> lm.fit <- lm(balance ~ gender, data=credit)
> summary(lm.fit)

Call:
lm(formula = balance ~ gender, data = credit)
Residuals:
    Min      1Q  Median      3Q     Max
-529.54 -455.35  -60.17  334.71 1489.20

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    509.80      33.13  15.389   <2e-16 ***
genderFemale    19.73      46.05   0.429    0.669
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 460.2 on 398 degrees of freedom
Multiple R-squared:  0.0004611,Adjusted R-squared:  -0.00205
F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

노트.

1. **회귀모형에서 가변수의 해석**

   genderFemale의 뜻은 Female일 때 이 가변수의 값이 1이라는 것이다.

   β_0 : 남성의 평균 신용카드 대출액

   $\beta_0 + \beta_1$: 여성의 평균 신용카드 대출액

2. **테이블의 해석** 절편의 p 값은 작지만 gender[Female]의 p 값은 크지 않다. 남성과 여성의 대출액의 차이가 크지 않다는 결론이다.

가변수(dummy variable) : 세 개 이상의 값을 갖는 변수

가변수의 정의

$$x_{i1} = \begin{cases} 1, & \text{i번째 사람이 동양인} \\ 0, & \text{i번째 사람이 동양인이 아닐 때} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{i번째 사람이 백인} \\ 0, & \text{i번째 사람이 백인이 아닐 때} \end{cases}$$

회귀모형에서 가변수의 해석

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \\
&= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{i번째 사람이 동양인} \\ \beta_0 + \beta_2 + \epsilon_i, & \text{i번째 사람이 백인} \\ \beta_0 + \epsilon_i, & \text{i번째 사람이 흑인} \end{cases}
\end{aligned}
$$

노트.

1. ethinicity는 백인, 흑인, 동양인 세 개의 값을 갖는다. 이를 두 개의 가변수를 이용해 정의할 수 있다.
2. **평균들의 해석**

   β_0 : 흑인의 평균 신용카드 대출액
   $\beta_0 + \beta_2$: 백인의 평균 신용카드 대출액
   $\beta_0 + \beta_1$: 동양인의 평균 신용카드 대출액

$$balance = \beta_0 + \beta_1 \times ethnicityAsian + \beta_2 \times ethnicityCaucasian$$

```
> lm.fit <- lm(balance ~ ethnicity, data=credit)
> summary(lm.fit)
Call:
lm(formula = balance ~ ethnicity, data = credit)
Residuals:
    Min      1Q  Median      3Q     Max
-531.00 -457.08  -63.25  339.25 1480.50

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          531.00      46.32  11.464   <2e-16 ***
ethnicityAsian       -18.69      65.02  -0.287    0.774
ethnicityCaucasian   -12.50      56.68  -0.221    0.826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188,Adjusted R-squared:  -0.004818
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

가변수 두 개가 포함된 회귀모형 추정치

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	-18.69	65.02	-0.287	0.7740
`ethnicity[Caucasian]`	-12.50	56.68	-0.221	0.8260

노트.

1. ethnicity[Asian]과 ethnicity[Caucasian]의 p 값이 크다.
   인종간의 차이가 크지 않다는 결론이다.

선형 모형의 핵심 가정들

1. **가법성 가정(additivity)** : 반응변수에 대한 예측변수 X_j의 효과는 다른 예측변수들의 값에 영향을 받지 않는다.
2. **선형성 가정(linearity)** : 예측변수의 변화에 따른 반응변수의 변화는 예측변수의 값에 영향을 받지 않는다.

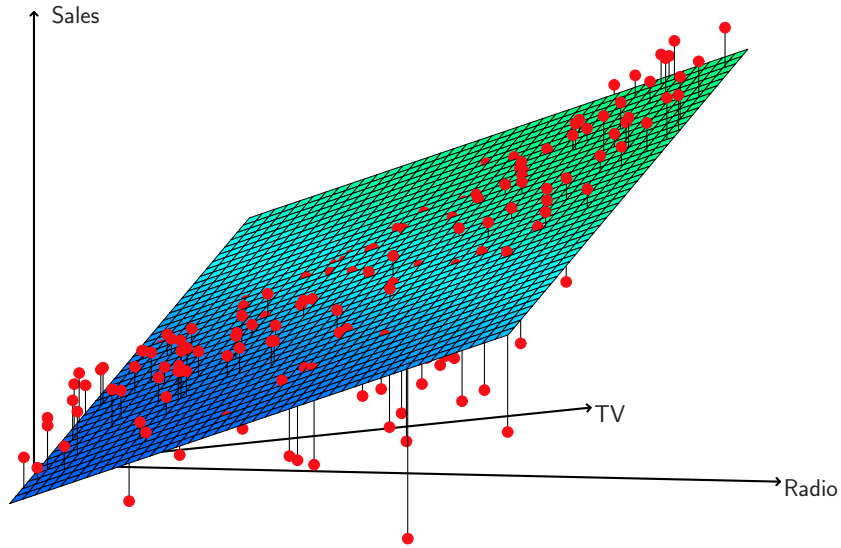
노트.

1. 이 두 가지 가정을 완화하거나 제거하는 방법에 대해 알아본다.
2. 가법성 가정이 깨지면 X_j가 Y에 미치는 영향이 다른 변수들의 값에 따라 달라진다.
3. 선형성 가정이 깨지면 X_j가 Y에 미치는 영향이 X_j의 값에 따라 달라진다.
4. 여기서 영향이라는 것은 회귀계수 β_j를 말한다. 즉, X_j의 단위 변화량에 대한 Y의 변화량.

가법성 가정의 완화 혹은 삭제

교호작용이 없는 모형

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$



교호작용이 있는 모형

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

노트.

1. 그림을 보면 라디오 광고비 지출을 늘리면 TV광고의 효과를 크게 한다. 라디오 광고비가 커질수록 TV 광고비의 기울기가 커진다.
2. 교호작용이 없는 모형은 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ X_1 이 한 단위 커질 때 Y는 β_1 만큼 커지고 이 값은 X_2의 값에 영향을 받지 않는다.
3. 교호작용이 있는 모형은 X_1의 기울기 $\beta_1 + \beta_3 X_2$는 X_2의 값에 따라 변하게 된다.

$balance = \beta_0 + \beta_1 \times tv + \beta_2 \times radio + \beta_3 \times tv \times radio$

```
> lm.fit = lm(sales ~ tv*radio, data=adv)
> summary(lm.fit)

Call:
lm(formula = sales ~ tv * radio, data = adv)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3366 -0.4028  0.1831  0.5948  1.5246

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
tv          1.910e-02  1.504e-03  12.699   <2e-16 ***
radio       2.886e-02  8.905e-03   3.241   0.0014 **
tv:radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,Adjusted R-squared:  0.9673
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

노트.

`sales ~ tv * radio` 대신 `sales ~ tv+radio+tv:radio`를 넣어도 동일한 결과를 얻는다.

교호작용이 포함된 추정치

$$
\begin{aligned}
\textit{sales} &= \beta_0 + \beta_1 \times \textit{TV} + \beta_2 \times \textit{radio} + \beta_3 \times \textit{radio} \times \textit{TV} + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_3 \times \textit{radio}) \times \textit{TV} + \beta_2 \times \textit{radio} + \epsilon \\
&= \beta_0 + \beta_1 \times \textit{TV} + (\beta_2 + \beta_3 \times \textit{TV}) \times \textit{radio} + \epsilon.
\end{aligned}
$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

계층 원칙(hierarchical principle)

교호작용항(X_1X_2)이 유의해서 모형에 포함시켜면 주효과 X_1 과 X_2 는 유의하지 않아도 모형에 포함시킨다.

노트.

1. 티비 광고를 1000불 증가시키면 매출이 $\beta_1 + \beta_3 \times radio$ 만큼 증가한다. 라디오 광고를 1000불 증가시키면 매출이 $\beta_2 + \beta_3 \times TV$ 만큼 증가한다.
2. **계층 원칙(hierarchical principle)** 교호작용항(X_1X_2)이 유의해서 모형에 포함시키면 주효과 X_1과 X_2는 유의하지 않아도 모형에 포함시킨다. X_1X_2가 유의하다면 X_1의 주효과가 정확히 0인지 아닌지는 관심이 없다.

가변수와 교호작용

교호작용이 없을 때

$$
\begin{aligned}
\textit{balance} &= \beta_0 + \beta_1 \times \textit{income} + \beta_2 \times \textit{student} + \epsilon \\
&= \begin{cases} (\beta_0 + \beta_2) + \beta_1 \times \textit{income} + \epsilon, & \text{student} = 1 \\ \beta_0 + \beta_1 \times \textit{income} + \epsilon, & \text{student} = 0 \end{cases}
\end{aligned}
$$

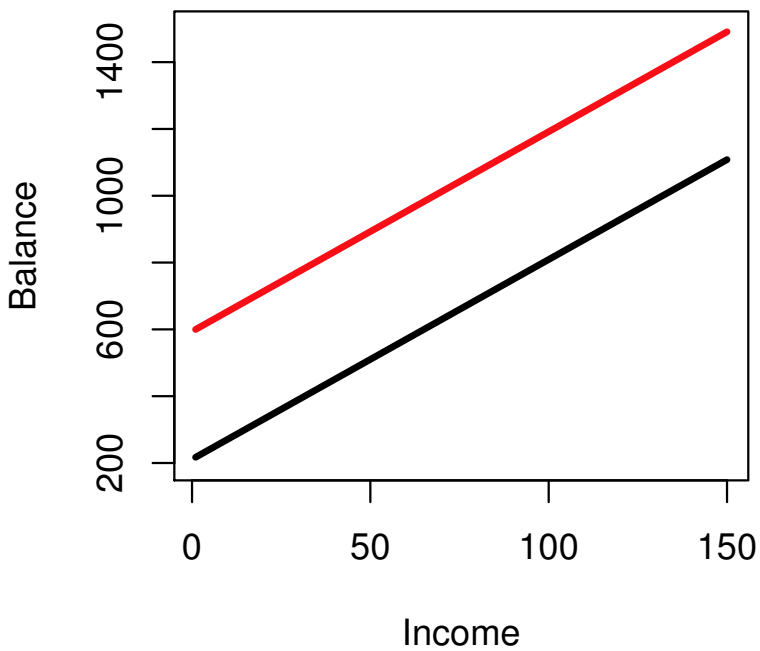
교호작용이 있을 때

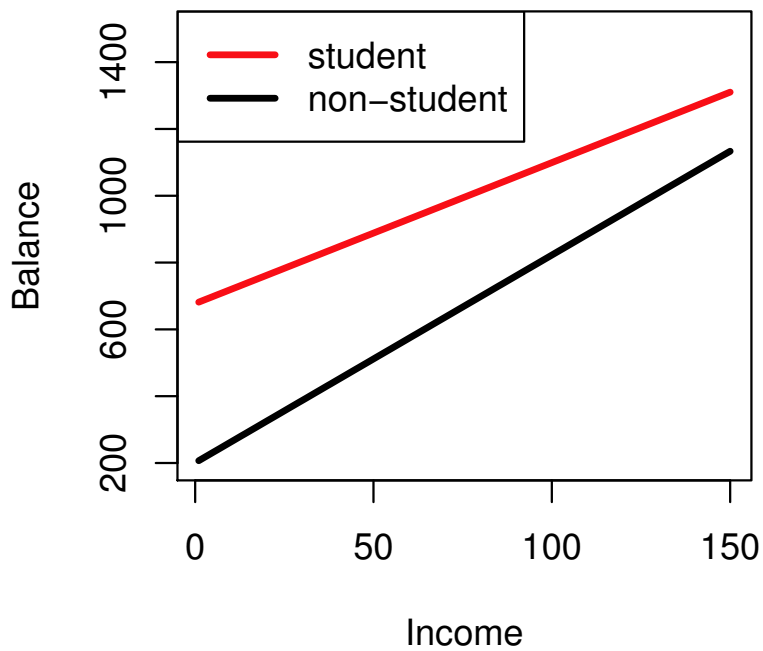
$$
\begin{aligned}
\textit{balance} &= \beta_0 + \beta_1 \times \textit{income} + \beta_2 \times \textit{student} + \beta_3 \times \textit{income} \times \textit{stu} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \textit{income} + \epsilon, & \text{student} = 1 \\ \beta_0 + \beta_1 \times \textit{income} + \epsilon, & \text{student} = 0 \end{cases}
\end{aligned}
$$

노트.

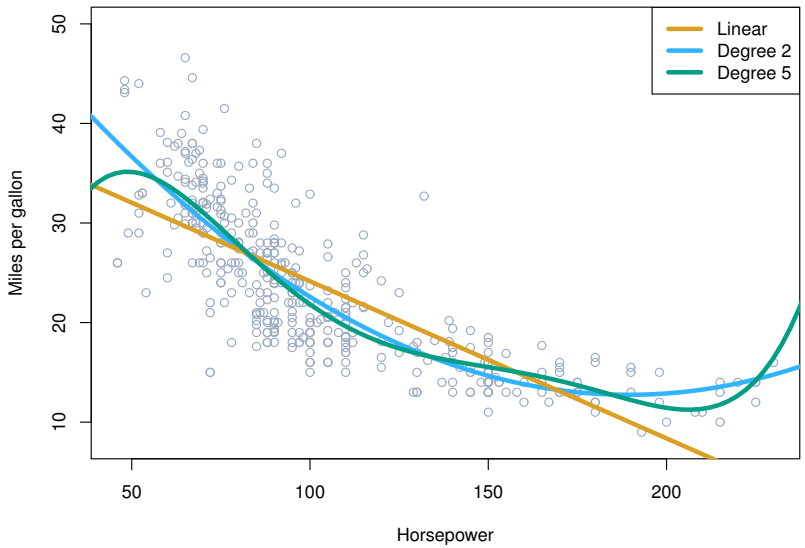
1. 교호작용이 없을 때는 절편이 각각 β_0와 $\beta_0 + \beta_2$인 두 개의 평행한 직선을 적합하는 것이다.
2. 교호작용이 있을 때는 두 개의 직선을 적합하는 것이다.
3. 다음 쪽의 그림은 이 상황을 나타낸다.

교호작용과 가변수


Balance
200
600
1000
1400
0
50
100
150
Income


student
non−student
Balance
200
600
1000
1400
0
50
100
150
Income

선형성 가정의 완화



	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

2차선형회귀모형

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon$$

노트.

1. 그림은 자동차 자료에서 horsepower와 mpg(miles per gallon)의 관계를 나타낸다. 오렌지 색은 1차 선형회귀직선을 적합한 결과이다. 자료는 비선형성을 보인다. 파란색은 다음의 2차식을 적합한 결과를 나타낸다.
2. $mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon$ 위의 모형은 horsepower의 함수로서는 비선형이지만 $horsepower$와 $horsepower^2$으로 봤을 때는 선형이다. 따라서 선형모형을 적합하는 방법으로 적합할 수 있다.
3. 그림에는 5차식을 적합한 결과도 있지만 너무 구불구불하다.

다항회귀 R 코드

```
> lm.fit = lm(mpg ~ poly(horsepower, 2, raw=T), data = auto)
> summary(lm.fit)

Call:
lm(formula = mpg ~ poly(horsepower, 2, raw = T), data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.9907  -6.0269  -0.2335   4.7160  23.8816

Coefficients:
                              Estimate Std. Error t value
(Intercept)                  16.8389456  0.9890343  17.026
poly(horsepower, 2, raw = T)1  0.1801992  0.0507205   3.553
poly(horsepower, 2, raw = T)2 -0.0007355  0.0005225  -1.408
                             Pr(>|t|)
(Intercept)                   < 2e-16 ***
poly(horsepower, 2, raw = T)1 0.000427 ***
poly(horsepower, 2, raw = T)2 0.160009
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.092 on 394 degrees of freedom
Multiple R-squared:  0.1829,Adjusted R-squared:  0.1787
```

노트.

1. 2차다항회귀를 적합한 결과이다.
2. `poly(horsepower, 2, raw=T)`는 `poly(horsepower, 2, raw=T)`를 의미한다. raw=F를 쓰면 orthogonal polynomial을 계산한다.
3. `I(horsepower^2)`는 계산한 식을 새로운 변수로 보라는 뜻이다. 다음과 같은 코드를 써도 동일한 결과를 얻는다.

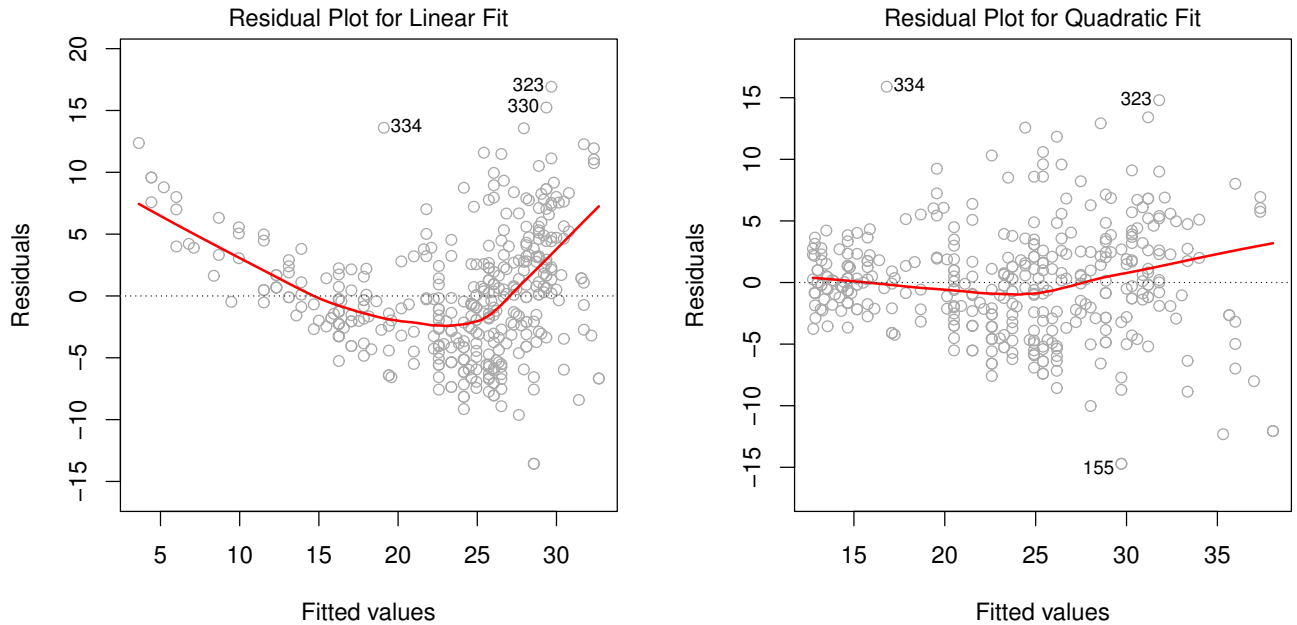
```
lm.fit = lm(mpg ~ horsepower + I(horsepower^2), data = auto)
summary(lm.fit)
```

노트. 선형모형을 적합할 때 발생할 수 있는 문제들. 이후에 다룰려는 내용들.

1. 예측변수-반응변수 관계의 비선형성
2. 오차항의 자기상관성
3. 오차의 이분산성(heteroscedacity)
4. 이상점(outlier)
5. 지렛대점(leverage point)
6. 다중공선성(multicolinearity)

이 문제들을 하나씩 다루려고 한다.

예측변수-반응변수 관계의 비선형성



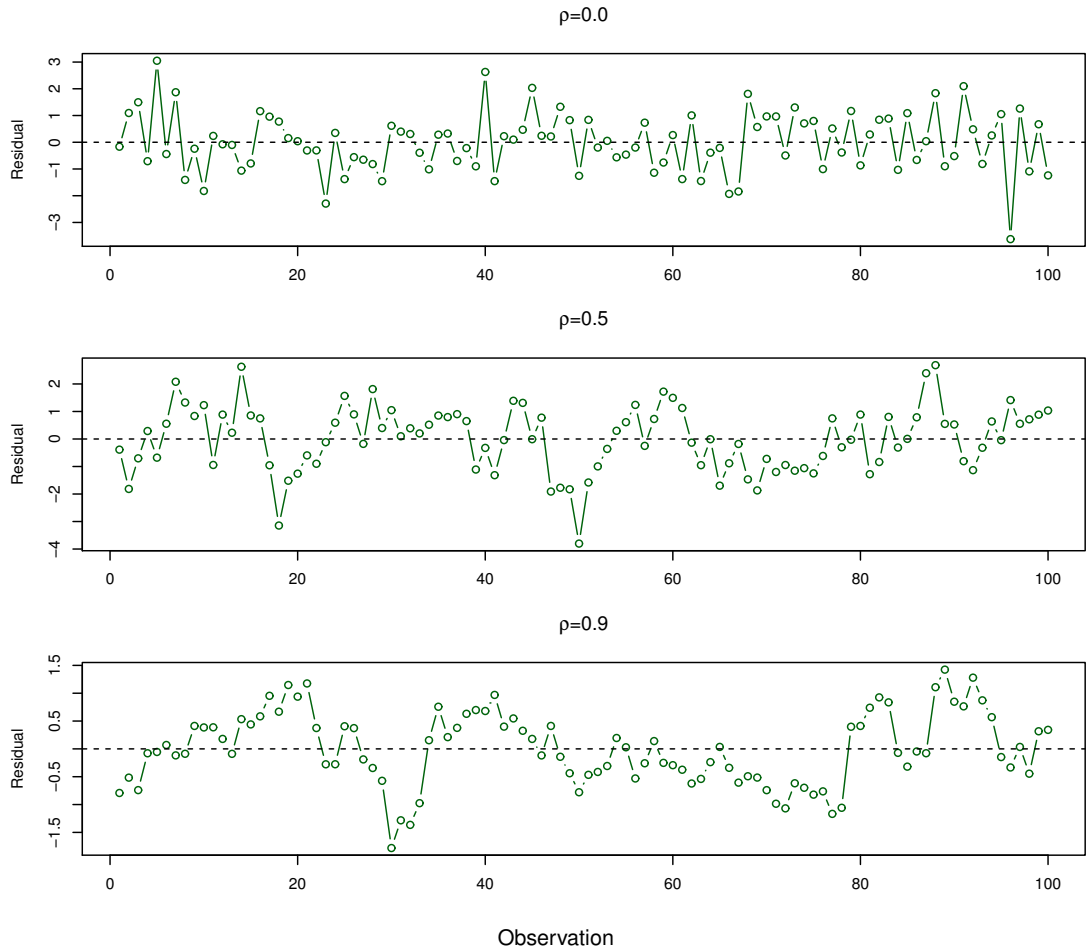
해결책

간단한 방법으로 문제가 되는 반응변수를 $\log X, \sqrt{X}, X^2$ 으로 변환하거나, 이 변수들을 포함하는 것이다. 다른 고급 방법들도 많이 존재한다.

노트. 예측변수-반응변수 관계의 비선형성

1. 예측변수-반응변수 관계의 비선형성은 예측값과 잔차의 그림을 통해서도 알수 있다.
2. 그림은 1차식을 적합했을 때와 2차식을 적합했을 때의 잔차그림을 보여준다. 잔차는 예측값의 값에 상관없이 0 근처이어야 하는데 1차식을 적합했을 때는 뚜렷한 패턴이 보인다. 2차식을 적합했을 때는 패턴이 많이 없어졌다.
3. 존재하는 고급방법들이란 비선형회귀나 비모수회귀모형을 말한다.

오차의 자기 상관성



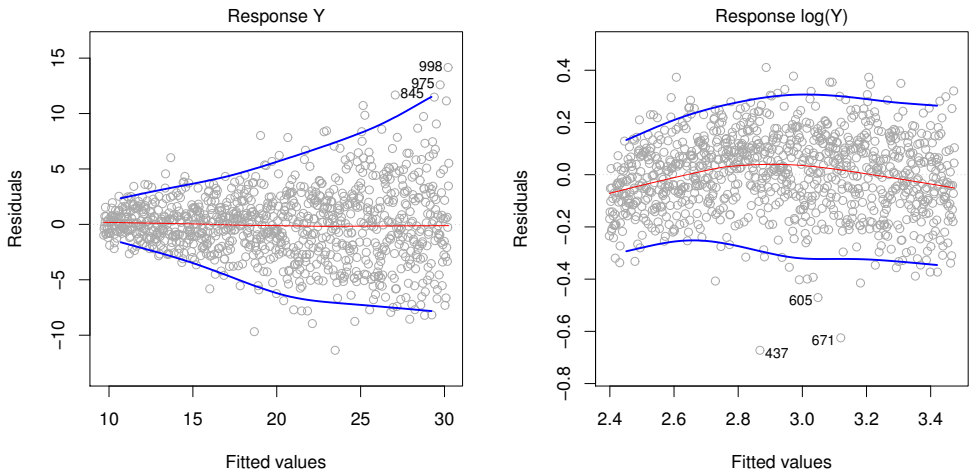
해결책

시계열 모형을 적용한다. 혹은 선형모형의 분산을 바꾼다.

노트.

1. 오차간에 상관성이 있는데 없다고 가정하고 모형을 적합하면 추정치들의 표준오차를 작게 추정하게 된다. 신뢰구간도 작게 구해진다. 시계열자료에 많이 나타난다.
2. $\rho = 0$일 때는 잔차그림에 패턴이 없다가 ρ가 커질수록 패턴이 점점 강하게 나타난다.

오차의 이분산성



해결책

1. Y를 $\log Y$나 $\sqrt{Y}$로 변환한다.
2. 가중최소제곱법(weighted least squares method)을 이용하여 추정한다.

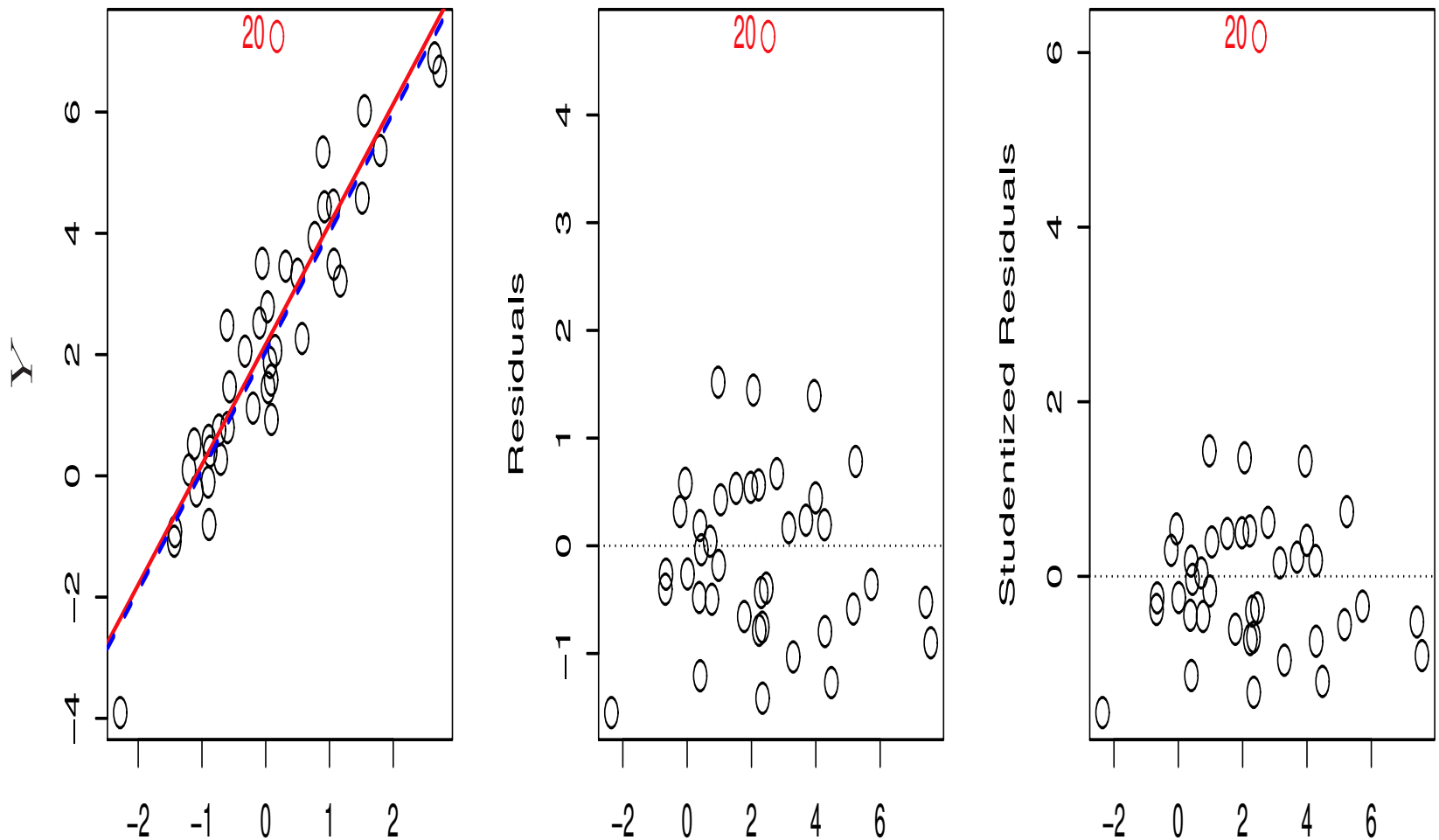
노트.

1. 등분상성 가정 $\mathbb{V}ar(\epsilon_i) = \sigma^2$ 이 성립하지 않는 경우이다
2. 각 관측치가 분산 σ^2 인 관측치 n_i 개의 평균일 때

$$\sigma_i^2 = \frac{\sigma^2}{n_i}$$

라는 것을 알 수 있다. 가중최소제곱법(weighted least squares method)을 이용하여 추정한다.

이상점



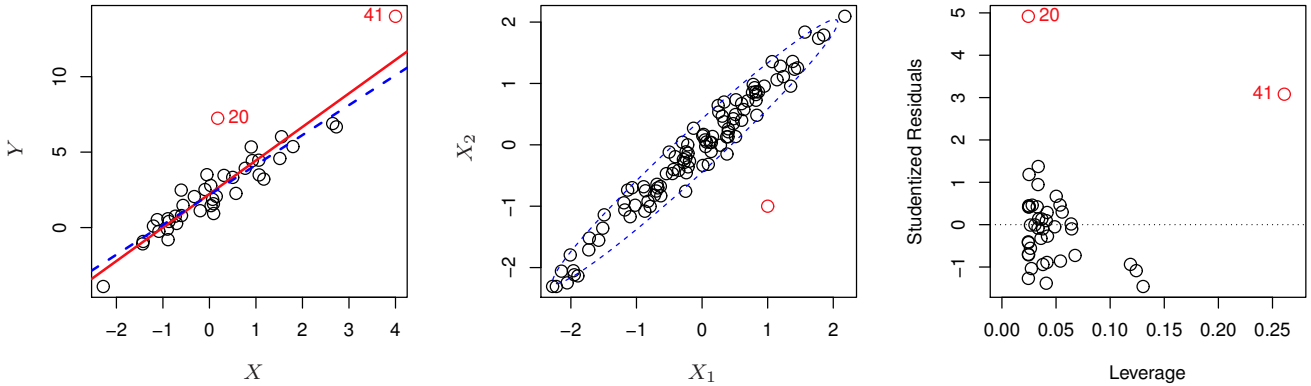
해결책

1. 이상점이 자료수집시 실수 때문에 발생한 것이면 제거가 가능하다. 그렇지 않을 경우는 중요한 예측변수가 포함이 안되었거나, 모형이 적합하지 않을 수 있을 수 있다. 이상점을 제거하는 것에 신중을 기해야한다.

노트.

1. 이상점이란 관측된 y 값과 예측값 $\hat{y}$ 이 많이 차이나는 경우를 말한다.
2. **영향**
   - 2.1 추정치는 큰 차이가 없을 수 있다. 앞 그림 에서 빨간선은 전체 자료를 다 포함한 회귀직선, 파란색 대시라인은 20번 관측치를 빼고 적합시킨 회귀직선이다. 큰 차이는 없어보인다.
   - 2.2 RSE$(1.09 \to 0.77)$에 큰 변화를 줄 수 있다. 이는 이후의 추론 즉 신뢰구간, 가설검정에 영향을 미칠 수 있다.
   - 2.3 $R^2(0.805 \to 0.892)$에 큰 변화를 줄 수 있다.

지렛대점



1. 보통의 범위에서 벗어난 x_i 값 때문에 추정식에 큰 영향을 미치는 점을 말한다.
2. 지렛대 통계량(leverage statistic)으로 체크한다. 단순회귀일 때는 다음과 같은 식을 갖는다.

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

노트.

1. 보통의 범위에서 벗어난 x_i 값 때문에 추정식에 큰 영향을 미치는 점을 말한다.
2. 그림을 보면, 관측치 20은 이상점이고 관측치 41번은 지렛대점이면서 동시에 이상점인 것을 알 수 있다.
3. 중앙의 그림은 X_1과 X_2 각 변수의 입장에서 보면 지렛대점이 아니나 두 변수를 동시에 그려보면 이상점이 되는 관측치를 보여준다. 따라서 중회귀분석에서는 지렛대점을 그림으로 확인하기가 어렵다.

1. 지렛대 통계량(leverage statistic)으로 체크한다.
   1.1 단순회귀일 때는 다음과 같은 식을 갖는다.

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

   1.2 중회귀모형에서 h_i 값은 $1/n \sim 1$ 사이의 값이다.

   1.3 전체 관측치의 평균은 $\frac{p+1}{n} = \frac{1}{n}\sum_{i=1}^{n} h_{ii}$ 이다. h_i 가 $\frac{p+1}{n}$ 보다 많이 크면(보통 $h_{ii} \geq \frac{3(p+1)}{n}$) 지렛대점이다.

노트. 지렛대 통계량

[1]선형모형

$$y = X\beta + \epsilon,\ \epsilon \sim (0, \sigma^2 I_n)$$

을 가정하자. 햇행렬(hat matrix)을 $H = X(X^TX)^{-1}X^T$ 라고 정의한다. 지렛대 통계량은 햇행렬의 i번째 대각원소

$$h_i = (H)_{ii}$$

로 정의된다.

성질들

1. $h_i = \dfrac{\partial \hat{y}_i}{\partial y_i}$
2. $0 \leq h_i \leq 1$
3. $\mathbb{V}ar(e_i) = (1 - h_i)\sigma^2$, $e_i = y_i - \hat{y}_i$.

의미

$\hat{y} = Hy$ 이다. 따라서 i번째 반응변수값 y_i의 예측치는

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{i1}y_1 + \ldots + \boldsymbol{h_{ii}y_i} + \ldots + h_{in}y_n$$

이 된다. 여기서 $H = (h_{ij})$로 나타낸다. 즉, $h_i = h_{ii}$는 $\hat{y}_i$에 y_i가 영향을 끼치는 정도이다. 이 값이 크면, 다른 반응변수 값들이 y_i를 예측하는 것과 자신이 예측하는 것이 많이 다르다는 것이다. $CV_{(n)}$ 공식을 보면

$$y_i - \hat{y}_i^* = \frac{y_i - \hat{y}_i}{1 - h_i}$$

노트. 참고. Cook's distance

3

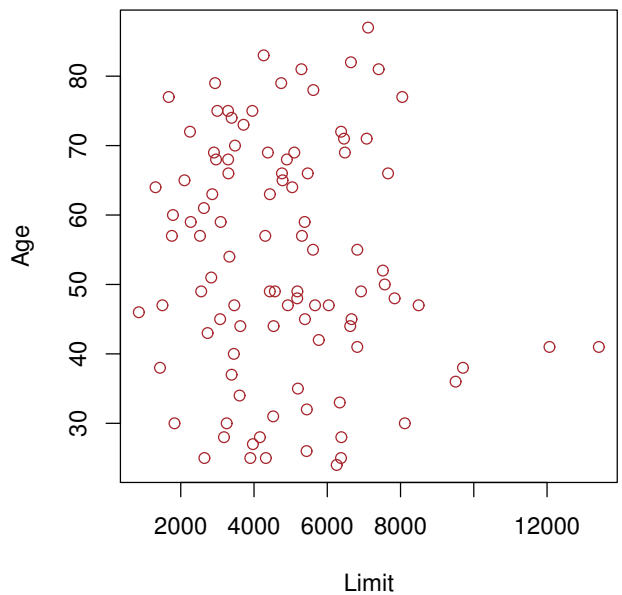
1. i번째 관측치가 influential point(leverage point와 같은 개념인 것 같다)인지 판단하는 기준 중 하나다.
2. i번째 관측치의 쿡 통계량 D_i는 관측치 i를 빼고 모형을 적합했을 때의 차이를 나타낸다. 그 정의는

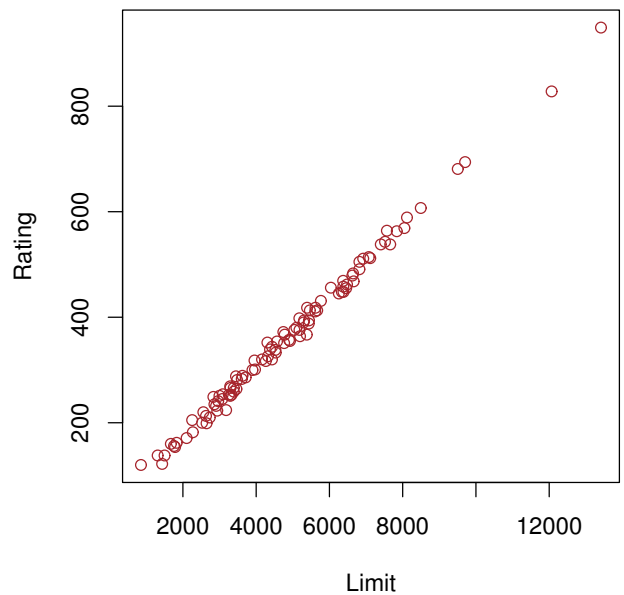
$$D_i = \frac{\sum_{i=1}^{n} (\hat{y}_i - \hat{y}_{j(-i)})^2}{(p+1)MSE}$$

   이다.
3. $D_i > 1$이면 크다고 한다(Cook and Weisberg 1982) 혹은 $D_i > \frac{4}{n}$이면 크다고 한다. (Bollen and Jackman 1990)
4. 이 외의 다른 개념들 DFFITS, Studentized residual.

[3] wikepedia를 참고했다.

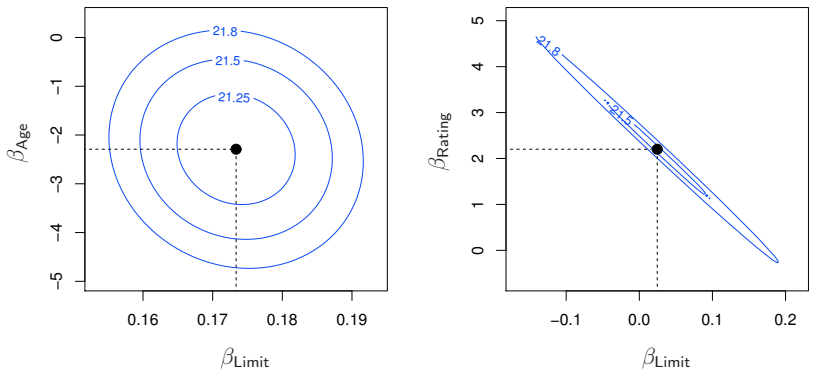
다중공선성





1. 두 개이상의 예측변수가 상관성이 큰 경우이다.

다중공선성이 야기하는 문제



1. 많은 $(\beta_{Limit}, \beta_{rating})$이 동일하게 자료를 설명한다.
2. 한 개의 예측변수만의 효과를 분리하기 어렵다.
3. 각 회귀계수 $(\beta_{Limit}, \beta_{rating})$의 변화 범위가 매우 크다.
4. 표준오차가 매우 커진다.
5. 통계적 가설 $H_0 : \beta = 0$ 의 검정력이 떨어진다.

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

노트.

1. 각 회귀계수 β_{Limit}, β_{rating})의 변화 범위가 매우 크다.
2. 표준오차가 매우 커진다. 테이블을 보면 모형 2의 $SE(\hat{\beta}_{Limit})$이 모형 1의 $SE(\hat{\beta}_{Limit})$ 보다 12배나 크다.
3. 표준오차가 커져서 t-통계량 값이 작아진다. 통계적 가설 $H_0 : \beta = 0$ 의 검정력이 떨어진다. 테이블을 보면 모형 1에서는 limit이 유의하게 나오지만 모형 2에서는 limit도 rating도 유의하지 않다.

다중공선성

감지(detect)하는 법

1. 예측 변수간의 상관계수를 본다.
2. 분산팽창인수(VIF, variance inflation factor)를 본다.

해결책

1. 문제가 되는 변수를 제거한다.
2. 문제가 되는 여러 개의 변수를 한 개의 변수로 합친다.

노트.

1. 하지만 모든 다중 공선성을 상관계수가 보여주지는 못한다. 두세개의 변수들이 동시에 한 변수와 상관성이 있을 때(이 때를 다중공선성이 있다고 한다.)는 상관계수로 알수가 없다. 이 때 분산팽창인수(VIF, variance inflation factor)를 본다.
2. i번째 예측변수의 분산팽창지수 $vif(X_j)$는

$$\frac{\text{현재의 } \mathbb{V}ar(\hat{\beta}_j)}{X_j\text{가 다른 예측변수들과 직교한다고 했을 때 } \mathbb{V}ar(\hat{\beta}_j)}$$

로 정의된다. 이 비율이다.

노트. 분산팽창지수

[4]

$$\hat{\mathbb{V}ar}(\hat{\beta}_j) = \frac{s^2}{(n-1)\hat{\mathbb{V}ar}(X_j)} \frac{1}{1-R_j^2}$$

와 같이 표현할 수 있다. 여기서

$s^2 = \sqrt{RSS/(n-p)}$

$\hat{\mathbb{V}ar}(X_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

$R_j^2 = X_j$를 반응변수로 나머지 변수들이 예측변수인 회귀모형의 결정계수

위의 식을 보면 $\frac{s^2}{(n-1)\hat{\mathbb{V}ar}(X_j)}$은 $X_j \perp X_{-j}$일 때 $\hat{\mathbb{V}ar}(\hat{\beta}_j)$이다. $\frac{1}{1-R_j^2}$를 X_j의 분산팽창지수라고 정의한다.

가장 작을 때는 1이고 클 때는 ∞까지도 간다.

[4] Wikipedia variance inflation factor를 보고 정리했다.

참고문헌

아래의 책에서 제공되는 그림을 써서 슬라이드를 만들었다.

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
An introduction to statistical learning.
Springer, 2013.