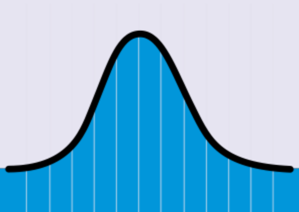
베이즈데이터분석 / 이재용 교수

13강

선형회귀모형

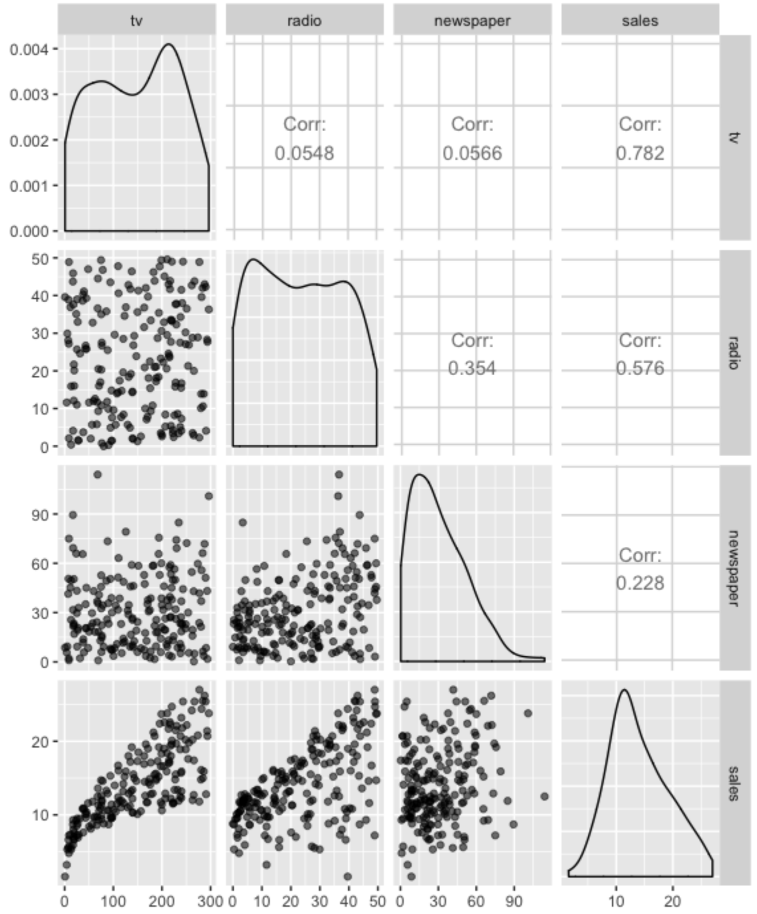
한국방송통신대학교

목차



목차

광고 (Advertising) 자료: 자료의 탐색



- 200개의 도시에서 한 해에 얻어진 매출, 텔레비전, 라디오, 신문 광고 자료이다.
- 매출의 단위는 천 개이고, 나머지의 단위는 천 불이다.
- 텔레비전, 라디오, 신문의 광고비용이 매출에 어떤 영향을 미치는가?

자료 탐색: 질문

1차원 자료 탐색

- 각 변수들의 요약 통계량을 구한다. summary
- 각 변수들의 히스토그램을 그린다. hist

2차원 자료 탐색

- 두 개의 변수 별로 2차원 산점도를 그린다. pairs, ggpairs
- 두 개의 변수간 상관 계수를 구한다. cor

자료 탐색: 질문

자료에 대한 질문

- 한 도시에서 1년에 보통 몇 개의 상품을 파나?
- 평균 TV 광고비는 얼마인가?
- 가장 많이 TV 광고비를 지출한 도시의 광고비 액수는?
  이 도시의 매출개수는?
- 매출이 가장 많은 도시의 매출은 얼마이고,
  이 도시의 TV, radio, newspaper 광고비는 얼마인가?
- 매출이 가장 작은 도시의 매출은 얼마이고,
  이 도시의 TV, radio, newspaper 광고비는 얼마인가?

두 변수의 관계에 관한 질문들

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

예측에 관한 질문

- 각 미디어의 매출에 대한 효과를 얼마나 정확히 추정할 수 있나?
  TV 광고에 1달에 만 불을 지출하면 매출이 얼마나 오르나?
- 미래의 매출액을 얼마나 정확히 예측할 수 있나?
  주어진 TV, 라디오, 신문 광고 비용을 지출하면
  매출이 얼마가 되나?

두 변수의 관계에 관한 질문들

관계에 관한 질문

- 광고 예산과 매출 사이에 관계가 있나? 없다면 광고를 하지 말자고 주장할 수 있다.
- 관계가 있다면 그 관계가 얼마나 강한가? 강한 관계가 있다면 사용된 광고비로 매출을 정확히 예측할 수 있고, 약한 관계라면 막연한 추측보다는 조금 더 나은 예측을 할 수 있을 것이다.
- 어떤 미디어가 매출에 영향을 미치나? 이 질문에 답하기 위해서는 각 미디어의 효과를 분리할 수 있어야 한다.
- 변수간 관계가 선형인가? 선형이면 선형 회귀를 이용하고, 아니면 다른 방법을 이용해야 한다.
- 광고 매체 사이에 시너지 효과가 있는가? 교호 작용(interaction effect)가 있는가?

회귀 분석은 위의 질문들에 대한 답변을 하는 과정이다.

회귀모형

$$y = f(x) + \epsilon$$

y : 반응 변수, 종속 변수

$x = (x_1, \ldots, x_p)$: 예측 변수, 독립 변수, 설명 변수

ϵ : 평균이 0인 랜덤 오차항

회귀 분석의 목적

- 예측(prediction): $\hat{y} = \hat{f}(x)$로 y값을 예측
- 추론(inference): x와 y의 관계를 이해

목차

단순 선형 회귀 모형(simple linear regression model)

모형

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0, β_1: 회귀 계수 ϵ: 오차항

β_1의 의미

x가 한 단위 커질 때,

늘어나는 $\mathbb{E}y$의 크기

예

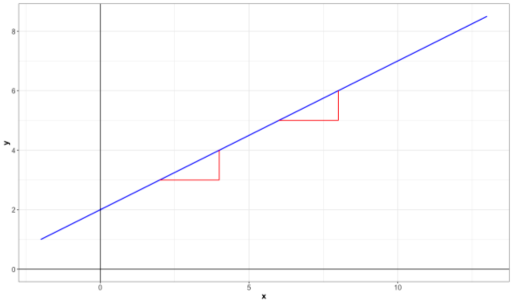
$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV}.$$

예측

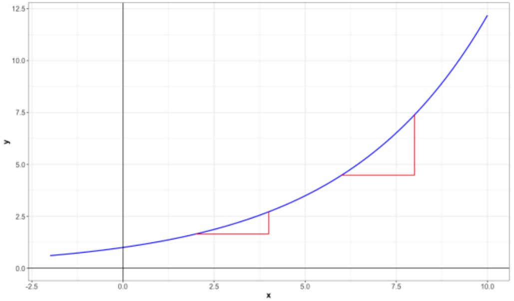
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

기울기 β_1의 의미

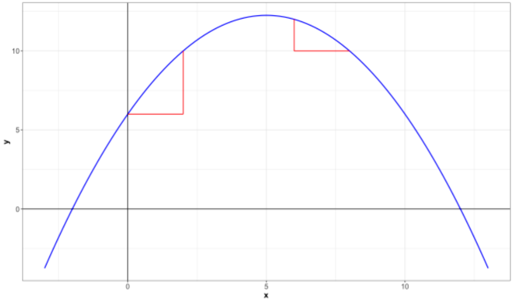
- 직선의 식. $y = \beta_0 + \beta_1 x$
- 직선의 특징은 기울기 $\beta_1 = \dfrac{y\text{의 변화량}}{x\text{의 변화량}}$이 일정하다



기울기 = 0.5와 0.5



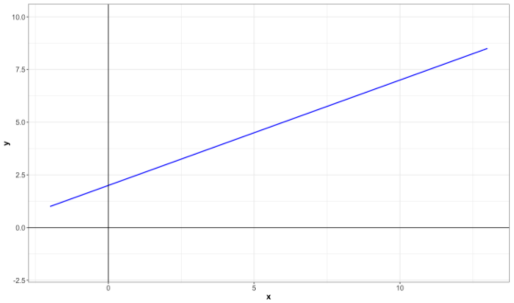
기울기= 0.53과 1.45



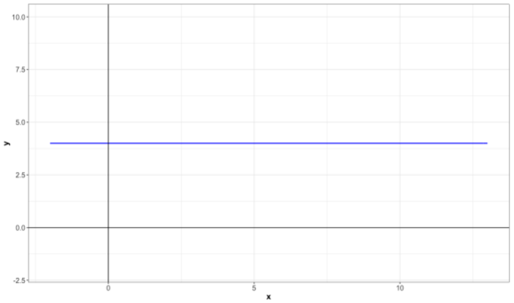
기울기= 2와 -1

기울기 β_1의 의미

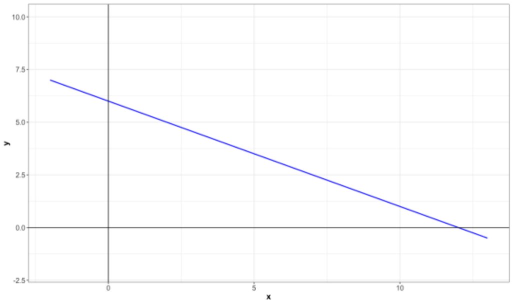
β_1의 값에 따른 직선의 변화



$\beta_1 > 0$



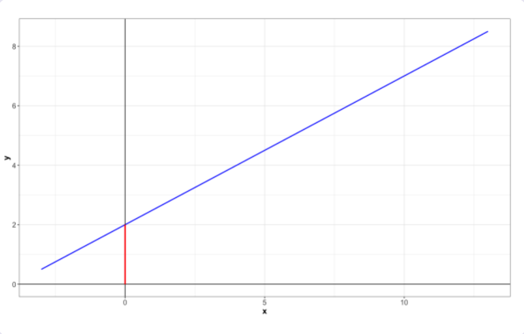
$\beta_1 = 0$



$\beta_1 < 0$

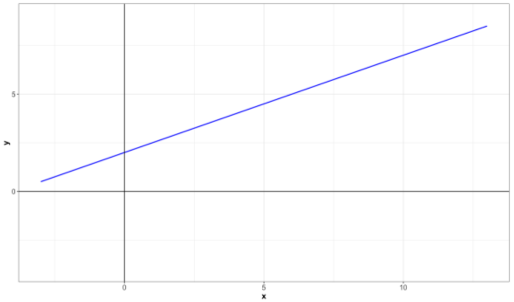
y 절편 β_0의 의미

- 직선의 식. $y = \beta_0 + \beta_1 x$
- $x = 0$일 때, y의 값. $y = \beta_0 + \beta_1 \times 0 = \beta_0$

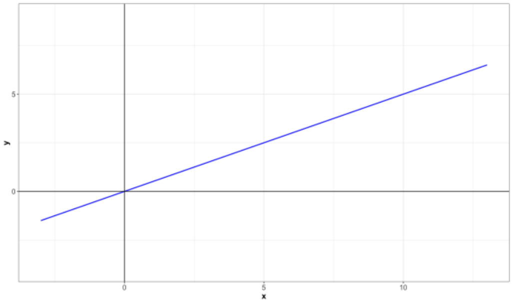


y 절편 β_0의 의미

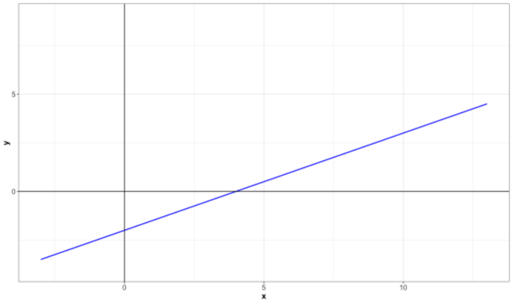
β_0의 값에 따른 직선의 변화



$\beta_0 > 0$



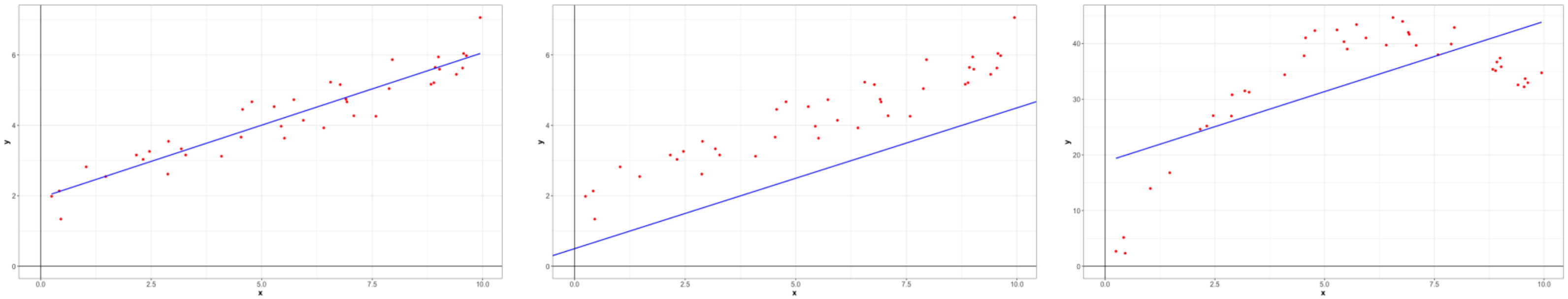
$\beta_0 = 0$



$\beta_0 < 0$

오차 항의 평균 $\mathbb{E}\epsilon$

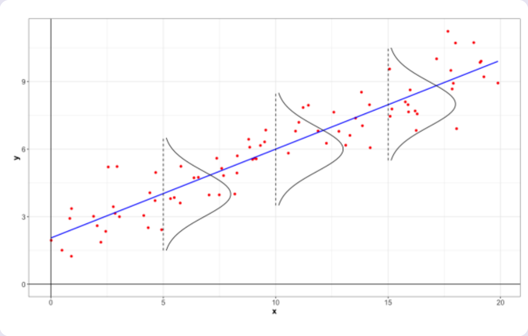
- 회귀식. $y = \beta_0 + \beta_1 x + \epsilon$.
- 오차항의 평균은 $\mathbb{E}\epsilon = 0$이고 분산은 $\mathbb{V}ar\epsilon = \sigma^2 > 0$
- $\mathbb{E}\epsilon = 0$이 아니라면.



$\mathbb{E}\epsilon = 0$ $\mathbb{E}\epsilon > 0$ $\mathbb{E}\epsilon$ 변동

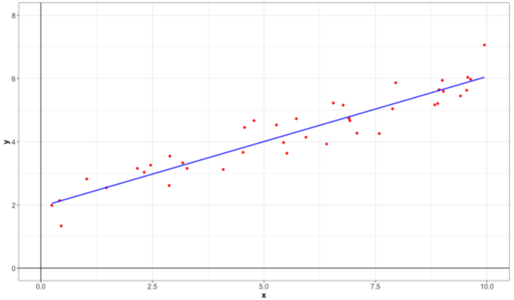
오차 항의 분산 $\mathbb{V}ar\epsilon = \sigma^2$

- 회귀식. $y = \beta_0 + \beta_1 x + \epsilon$.
- $\sigma^2 = \mathbb{V}ar\epsilon$는 고정된 x 값에서 y의 분산

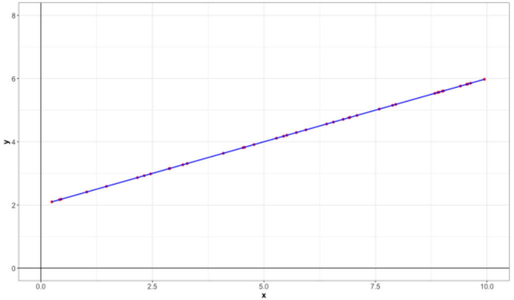


오차 항의 분산 $\mathbb{V}ar\epsilon = \sigma^2$

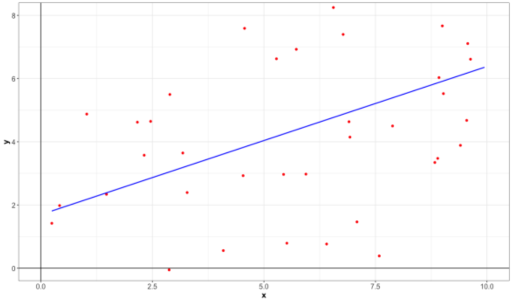
- $\sigma^2 = \mathbb{V}ar\epsilon$의 값에 따른 데이터 분포의 변화



σ^2작을 때



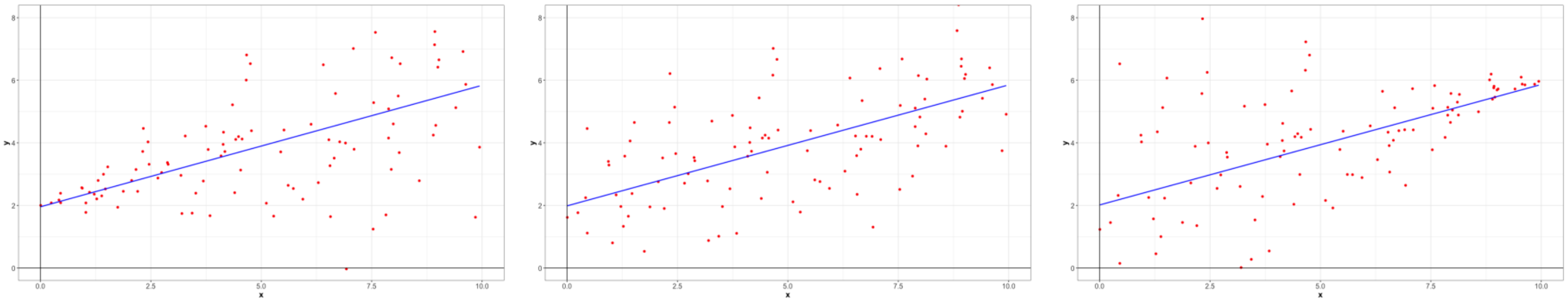
$\sigma^2 = 0$



σ^2클 때

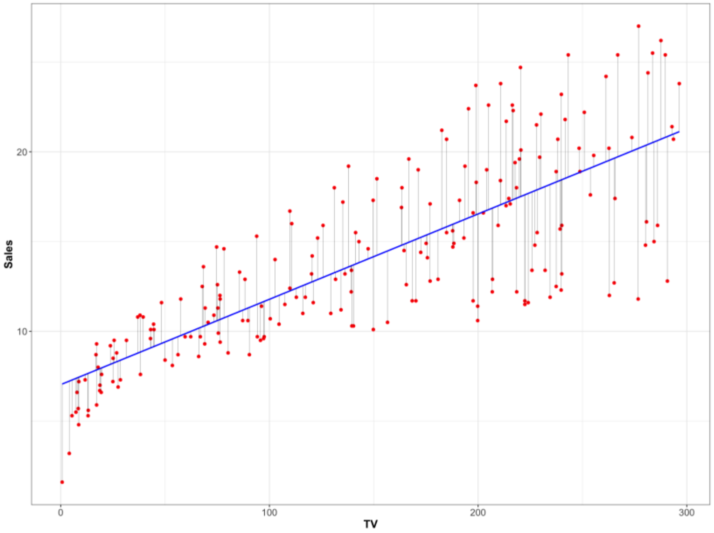
오차 항의 분산 $\mathbb{V}ar\epsilon = \sigma^2$은 x 값에 불변

- 회귀식. $y = \beta_0 + \beta_1 x + \epsilon.$
- $\sigma^2 = \mathbb{V}ar\epsilon$는 고정된 x 값에서 y의 분산
- $\sigma^2 = \mathbb{V}ar\epsilon$이 x 값에 따라 변한다면.

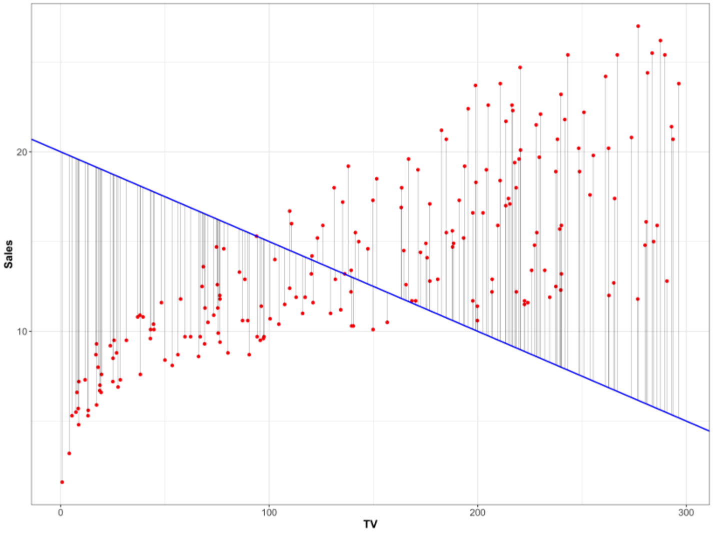


σ^2 커질 때　　σ^2 동일할 때　　σ^2작아질 때

회귀계수의 추정



$$y = 7.03259 + 0.04754x$$



$$y = 20.0 - 0.05x$$

어떤 직선이 데이터를 더 잘 설명하나?

- (모형과 사전분포)

$$sales_i = \beta_0 + \beta_1 * tv_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

$$\pi(\beta_0, \beta_1, \sigma^2) d\beta_0 d\beta_1 d\sigma^2 = \frac{1}{\sigma^2} d\beta_0 d\beta_1 d\sigma^2$$

- 판매량의 예측식을 써보자.

- 텔레비전 광고가 $147,000일 때, 판매량의 예측값은?

광고 자료 실습

코드

```
adv.lm1.code ="
data {
    int<lower=0> n;
    vector[n] tv;
    vector[n] sales;
}

parameters {
    real beta0;
    real beta1;
    real<lower=0> sigma;
}
model {
    for(i in 1:n) {
      sales[i] ~ normal(beta0 + beta1*tv[i], sigma);
    }
    target += 1/sigma^2;
}
"

data=list(n=dim(adv)[1], sales=adv$sales, tv=adv$tv)

adv.lm1 = stan(model_code=adv.lm1.code, data=data,
                        seed=1234567,chains=4, iter=5000, thin=1)
print(adv.lm1)
plot(adv.lm1, plotfun="dens")
plot(adv.lm1, plotfun="trace")
plot(adv.lm1, plotfun="ac")
```

목차

- 회귀 모형의 소개 및 목적
- 단순 선형 회귀 모형
- 중회귀모형

중회귀 모형

모형

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

x_j: j번째 예측 변수

β_j: x_j의 회귀 계수

ϵ : 오차

중회귀 모형에서 β_1의 의미

$x_2, \ldots, x_p$값이 고정되었을 때,

x_1이 한 단위 커질 때 커지는 $\mathbb{E}y$의 크기

중회귀 모형

예

$$sales = \beta_0 + \beta_1 \times tv$$

$$+ \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

예측식

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

중회귀 모형에서 β_j의 의미

- 변수가 두 개만 있는 경우를 고려하자.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- (β_1의 의미) x_2값이 고정되었을 때,
  x_1이 한 단위 커질 때 커지는 $\mathbb{E}y$의 크기.

- 일반적인 고정된 x_2에 대해,

$$y = (\beta_0 + \beta_2 x_2) + \beta_1 x_1.$$

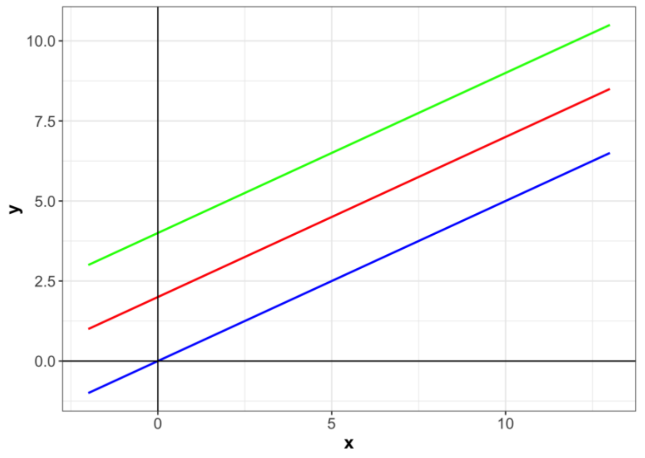
  $(\beta_0 + \beta_2 x_2)$는 절편이 되고,
  β_1은 x_1의 기울기가 된다.

중회귀 모형에서 β_j의 의미

- $x_2 = 0, 1, 1.2, 1.5, 2,$이어도,

$$\beta_1 = \frac{y\text{의 변화량}}{x_1\text{의 변화량}}$$

으로 동일하다.



파란색 직선은 $x_2 = -1,$

빨간색은 $x_2 = 0,$

초록색은 $x_2 = 1$일 때.

모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, 2, \ldots, n$$

p는 변수의 개수이고, $q = p + 1$이다.

행렬식을 이용한 모형

$$y_{n\times1} = X_{n\times q}\beta_{q\times1} + \epsilon_{n\times1}, \epsilon \sim N(0, \sigma I_n)$$

$$y = (y_1, \ldots, y_n)^T$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)$$

$$\epsilon = (\epsilon_1, \ldots, \epsilon_n)$$

- (모형과 사전분포)

$$sales_i = \beta_0 + \beta_1 * tv_i + \beta_2 * radio_i +$$

$$\beta_3 * newspaper_i + \epsilon, \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

$$\pi(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2) d\beta_0 d\beta_1 d\beta_2 d\beta_3 d\sigma^2$$

$$= \frac{1}{\sigma^2} d\beta_0 d\beta_1 d\beta_2 d\beta_3 d\sigma^2$$

광고 자료 실습

- 판매량의 예측식을 써보자.
- 티비, 라디오, 신문 광고가 각각 $147,000, $100,000, $50,000일 때, 판매량의 예측값은?

```
print(adv.lm2)
plot(adv.lm2, plotfun="dens")
plot(adv.lm2, plotfun="trace")
plot(adv.lm2, plotfun="ac")
```

코드

```
adv.lm2.code ="
data {
  int<lower=0> n;
  vector[n] tv;
  vector[n] radio;
  vector[n] newspaper;
  vector[n] sales;
}

parameters {
  real beta0;
  real beta1;
  real beta2;
  real beta3;
  real<lower=0> sigma;
}
```

코드

```
model {
  for(i in 1:n) {
    sales[i] ~ normal(beta0 + beta1*tv[i] + beta2*radio[i] +
                                          beta3*newspaper[i], sigma);
  }

  target += 1/sigma^2;
}
"

data=list(n=dim(adv)[1], sales=adv$sales, tv=adv$tv,
                radio=adv$radio, newspaper=adv$newspaper)

adv.lm2 = stan(model_code=adv.lm2.code, data=data,
                    seed=1234567,   chains=4,    iter=5000,    thin=1)
```

다음시간

14강

일반화 선형모형