Ch.3 Linear Regression

김철

YBigTa Data Science Team (ML Class)

목차

Linear Regression 이란?

- 선형 회귀분석은 연속형 종속변수 Y 와 독립변수들 $x_1,\ x_2,\ \cdots,\ x_p$ 사이의 **선형관계**를 적합시키기 위하여 사용된다.

Linear Regression 이란?

- 선형 회귀분석은 연속형 종속변수 Y 와 독립변수들 $x_1,\ x_2,\ \cdots,\ x_p$ 사이의 **선형관계**를 적합시키기 위하여 사용된다.
- 다중 선형 회귀모형은 모집단에 대하여 다음과 같은 관계를 가정한다.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Linear Regression 이란?

- 선형 회귀분석은 연속형 종속변수 Y 와 독립변수들 $x_1,\ x_2,\ \cdots,\ x_p$ 사이의 **선형관계**를 적합시키기 위하여 사용된다.
- 다중 선형 회귀모형은 모집단에 대하여 다음과 같은 관계를 가정한다.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

- 여기서 $\beta_0,\ \cdots,\ \beta_p$ 는 회귀계수를 의미하고, ϵ 는 잡음(noise)으로서 모형에 의해 설명되지 않는 부분을 의미한다. 모형의 회귀계수를 추정하기 위하여 모집단으로부터 추출된 표본 데이터가 사용된다.

'추론(Inference)'과 '예측(Prediction)'의 관점

- **추론(Inference)**: 선형회귀에서 추론이란 독립변수 $x_1, x_2, \cdots, x_p$ 의 선형함수로써 Y 가 어떻게 변하는지 이해하고자 하는 것. 좀더 상세하게는,

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

를 추정한

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

를 이용하여 모집단에 대한 추론을 수행함.

'추론(Inference)'과 '예측(Prediction)'의 관점

- **예측(Prediction)**: 독립변수들의 표본 데이터가 아닌 새로운 데이터로 Y의 실제값을 예측하는 것.

'추론(Inference)'과 '예측(Prediction)'의 관점

- **예측(Prediction)**: 독립변수들의 표본 데이터가 아닌 새로운 데이터로 Y의 실제값을 예측하는 것.
- 추정된 회귀모형 $\hat{Y}$이 Y에 대한 정확한 예측을 제공하기만 한다면, 개별 회귀계수들이 유의하지 않더라도 OK.

'추론(Inference)'과 '예측(Prediction)'의 관점

- 추론(Inference)과 예측(Prediction)의 예

'추론(Inference)'과 '예측(Prediction)'의 관점

- 추론(Inference)과 예측(Prediction)의 예
- 예) 부동산 시장에서 범죄율(x_1), 지역(x_2), 강과의 거리(x_3), 공기의 청정도(x_4), 학교(x_5), 지역의 소득 수준(x_6), 집의 크기(x_7) 등과 같은 독립변수들과 집값(y)을 연관시키고자 할 수 있다.

'추론(Inference)'과 '예측(Prediction)'의 관점

- 추론(Inference)과 예측(Prediction)의 예
- 예) 부동산 시장에서 범죄율(x_1), 지역(x_2), 강과의 거리(x_3), 공기의 청정도(x_4), 학교(x_5), 지역의 소득 수준(x_6), 집의 크기(x_7) 등과 같은 독립변수들과 집값(y)을 연관시키고자 할 수 있다.
- 이 때, 개별 독립 변수들이 집값에 어떻게 영향을 미치는가?

'추론(Inference)'과 '예측(Prediction)'의 관점

- 추론(Inference)과 예측(Prediction)의 예
- 예) 부동산 시장에서 범죄율(x_1), 지역(x_2), 강과의 거리(x_3), 공기의 청정도(x_4), 학교(x_5), 지역의 소득 수준(x_6), 집의 크기(x_7) 등과 같은 독립변수들과 집값(y)을 연관시키고자 할 수 있다.
- 이 때, 개별 독립 변수들이 집값에 어떻게 영향을 미치는가?
- 즉, "만약 집의 전망이 강을 내려다 볼 수 있다면 그 집의 가치가 얼마나 더 올라가는가?" $\longrightarrow$ **추론 문제**

'추론(Inference)'과 '예측(Prediction)'의 관점

- 추론(Inference)과 예측(Prediction)의 예
- 예) 부동산 시장에서 범죄율(x_1), 지역(x_2), 강과의 거리(x_3), 공기의 청정도(x_4), 학교(x_5), 지역의 소득 수준(x_6), 집의 크기(x_7) 등과 같은 독립변수들과 집값(y)을 연관시키고자 할 수 있다.
- 이 때, 개별 독립 변수들이 집값에 어떻게 영향을 미치는가?
- 즉, "만약 집의 전망이 강을 내려다 볼 수 있다면 그 집의 가치가 얼마나 더 올라가는가?" $\longrightarrow$ **추론 문제**
- 아니면 단순히, "주어진 집의 특징들에 대해 그 집의 가치를 예측하고 싶다." $\longrightarrow$ **예측 문제**

'추론(Inference)'과 '예측(Prediction)'의 관점

	통계학	**머신러닝**
회귀분석의 목적	추론	예측
모형적합시 데이터	전체 데이터 세트를 사용	training/ validation 데이터 세트로 나눔
성능측정의 기준	데이터가 모형에 얼마나 잘 적합하는지 예) R^2, $F-$검정	예측의 정확도 예) MSE

우리는 왜 Linear Regression을 사용하는가?

- Linear Regression의 기본 가정을 만족하는 모형은 높은 **설명력**과 예측력을 기대할 수 있다.

우리는 왜 Linear Regression을 사용하는가?

- Linear Regression의 기본 가정을 만족하는 모형은 높은 **설명력**과 예측력을 기대할 수 있다.
- '설명력이 좋다?!'

우리는 왜 Linear Regression을 사용하는가?

- Linear Regression의 기본 가정을 만족하는 모형은 높은 **설명력**과 예측력을 기대할 수 있다.
- '설명력이 좋다?!'
  - 종속변수와 독립변수들간의 관계 설명 가능

우리는 왜 Linear Regression을 사용하는가?

- Linear Regression의 기본 가정을 만족하는 모형은 높은 **설명력**과 예측력을 기대할 수 있다.
- '설명력이 좋다?!'
  - 종속변수와 독립변수들간의 관계 설명 가능
  - 내가 **예측한 값이 틀렸을 가능성**까지도 제시 가능!!!! So Amazing!!!

미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

- 미아리 점쟁이 아줌마와 배우신 분(Data Scientist)의 예측대결

미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

- 미아리 점쟁이 아줌마와 배우신 분(Data Scientist)의 예측대결
- 점쟁이 아줌마의 무기

미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

- 미아리 점쟁이 아줌마와 배우신 분(Data Scientist)의 예측대결
- 점쟁이 아줌마의 무기
  - 그녀의 감 or 구슬의 목소리

미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

- 미아리 점쟁이 아줌마와 배우신 분(Data Scientist)의 예측대결
- 점쟁이 아줌마의 무기
  - 그녀의 감 or 구슬의 목소리
- 배우신 분의 무기

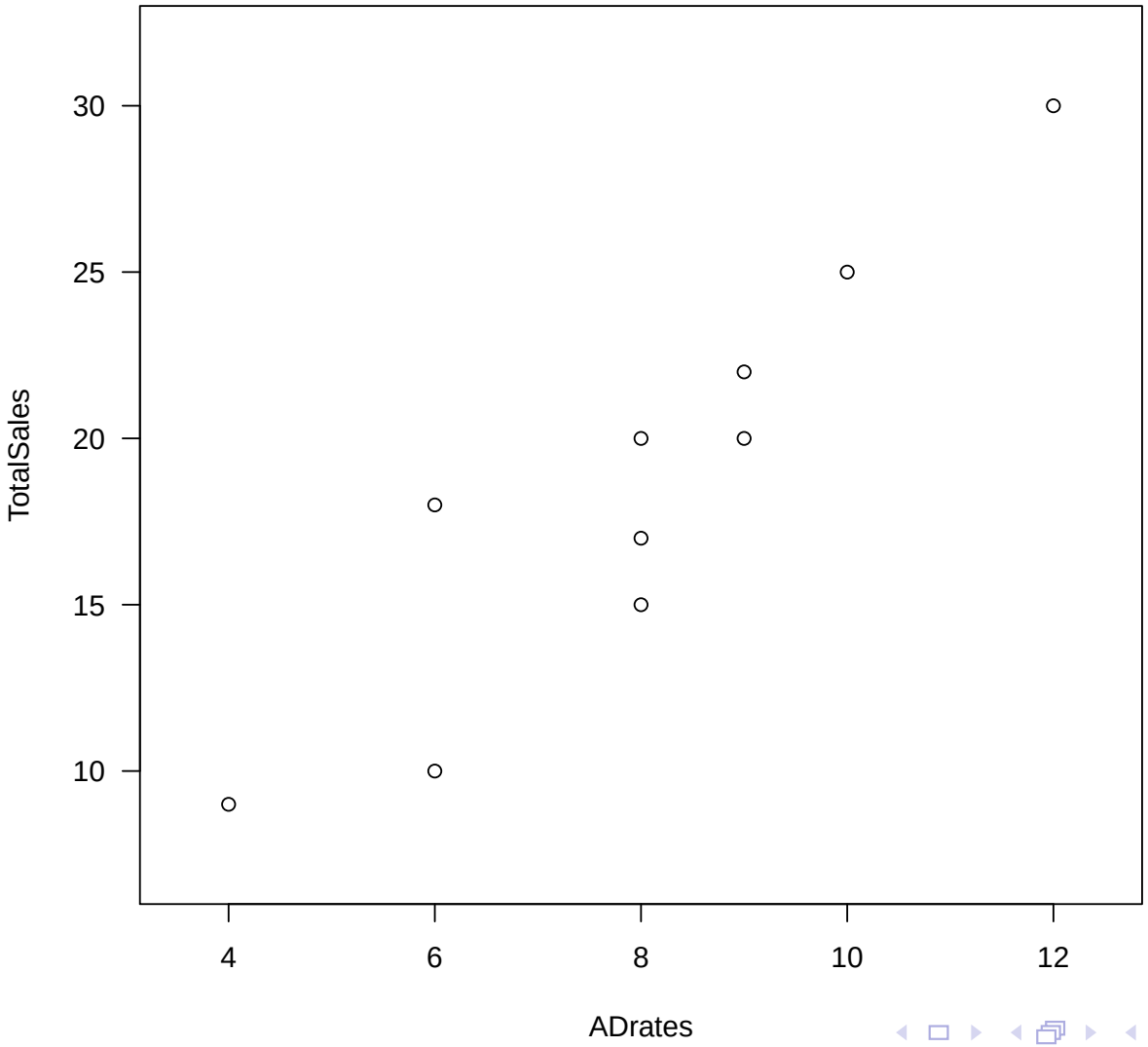
미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

- 미아리 점쟁이 아줌마와 배우신 분(Data Scientist)의 예측대결
- 점쟁이 아줌마의 무기
  - 그녀의 감 or 구슬의 목소리
- 배우신 분의 무기
  - 킹갓 The **'Linear Regression model'**

미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

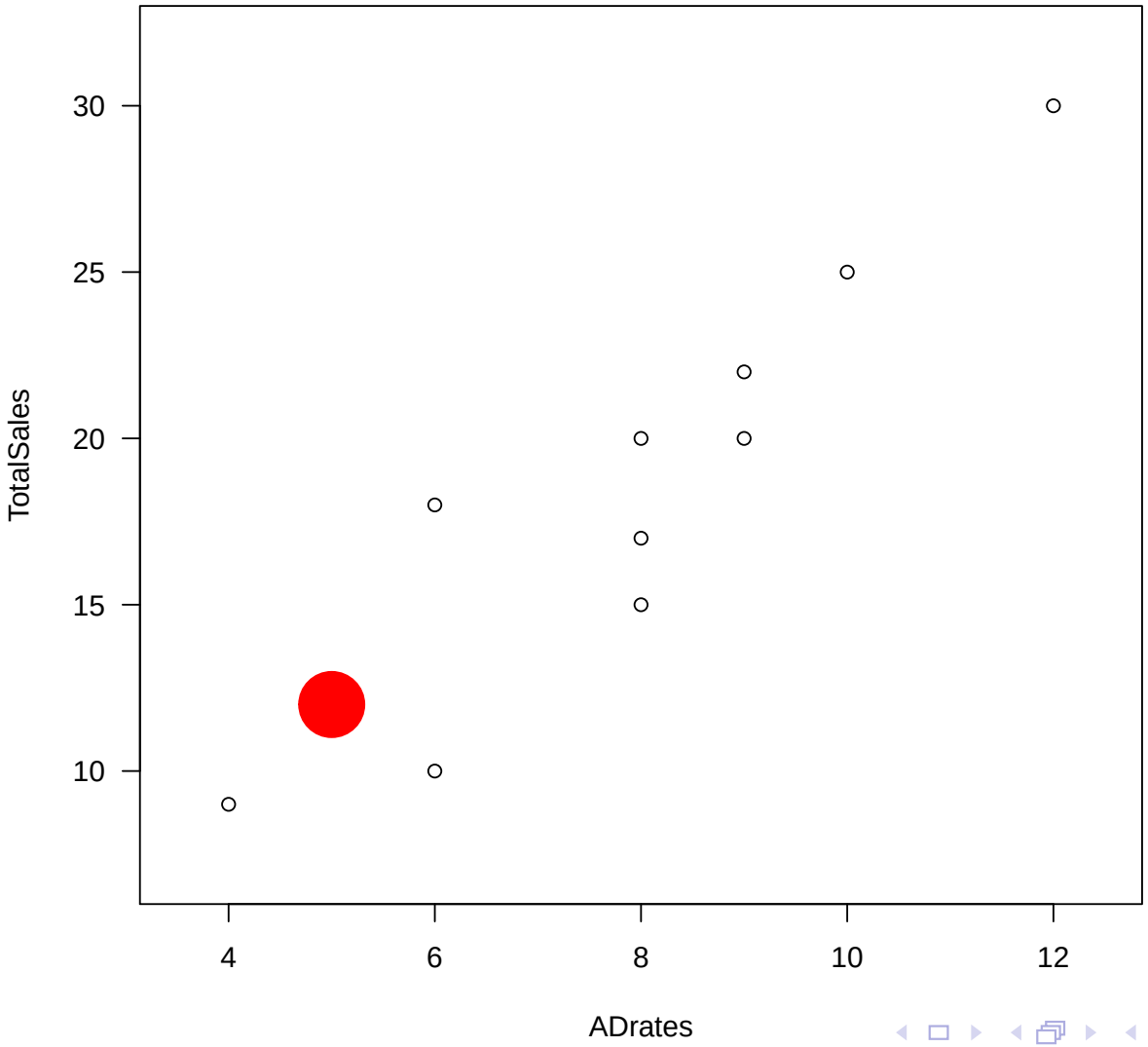
상점 번호	광고료 (단위 : 10 만 원)	총판매액 (단위: 100 만 원)	상점 번호	광고료 (단위 : 10 만 원)	총판매액 (단위: 100 만 원)
1	4	9	**6**	12	30
2	8	20	**7**	6	18
3	9	22	**8**	10	25
4	8	15	**9**	6	10
5	8	17	**10**	9	20

미아리 점쟁이 아줌마 vs 배우신 분(Data Scientist)

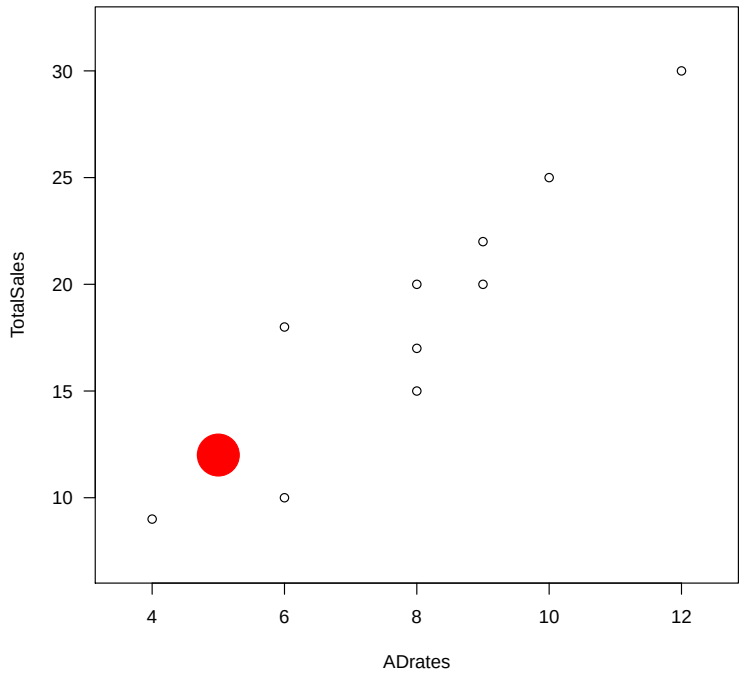


미아리 점쟁이 아줌마의 예측

미아리 점쟁이 아줌마의 예측

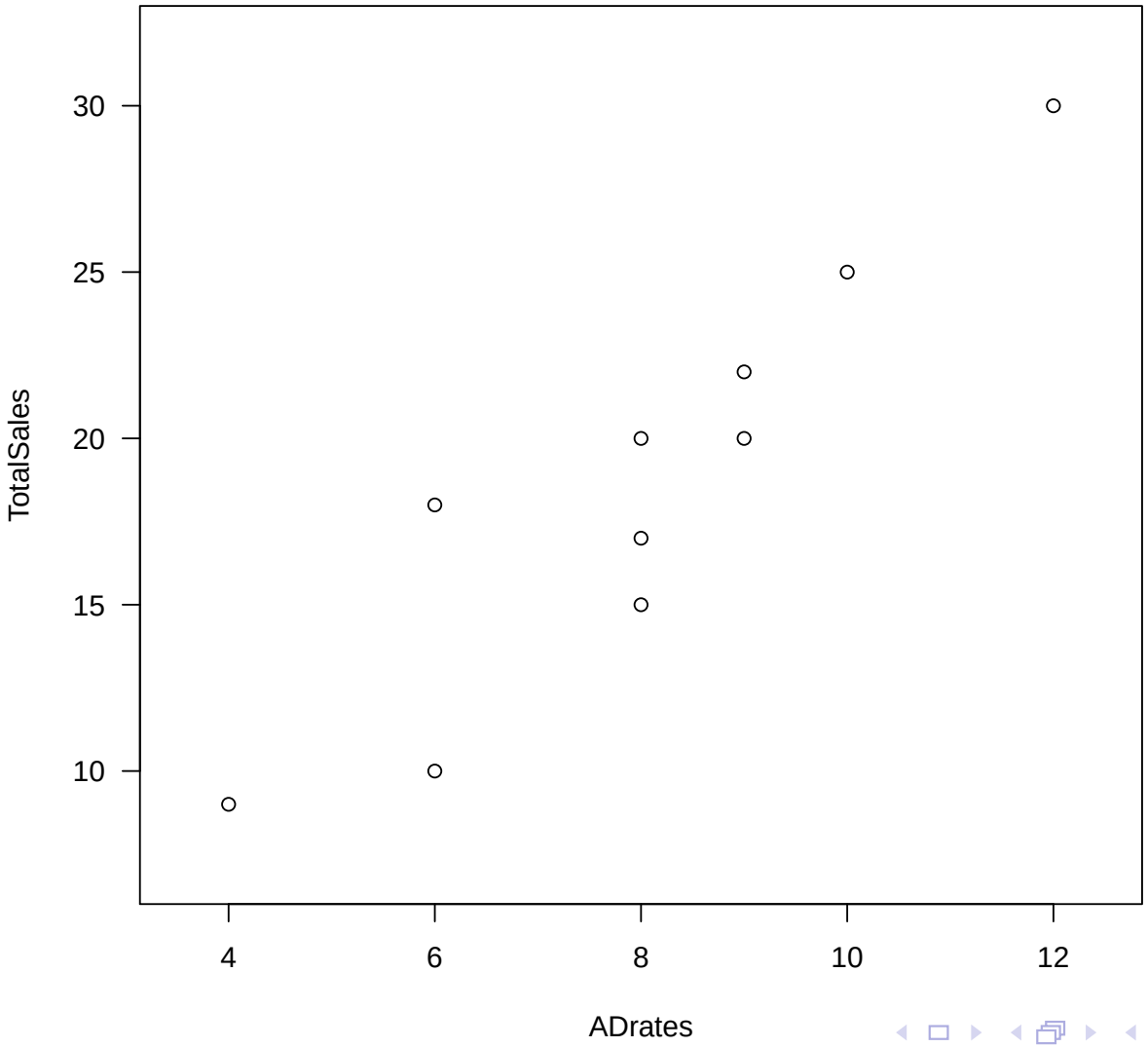


미아리 점쟁이 아줌마의 예측



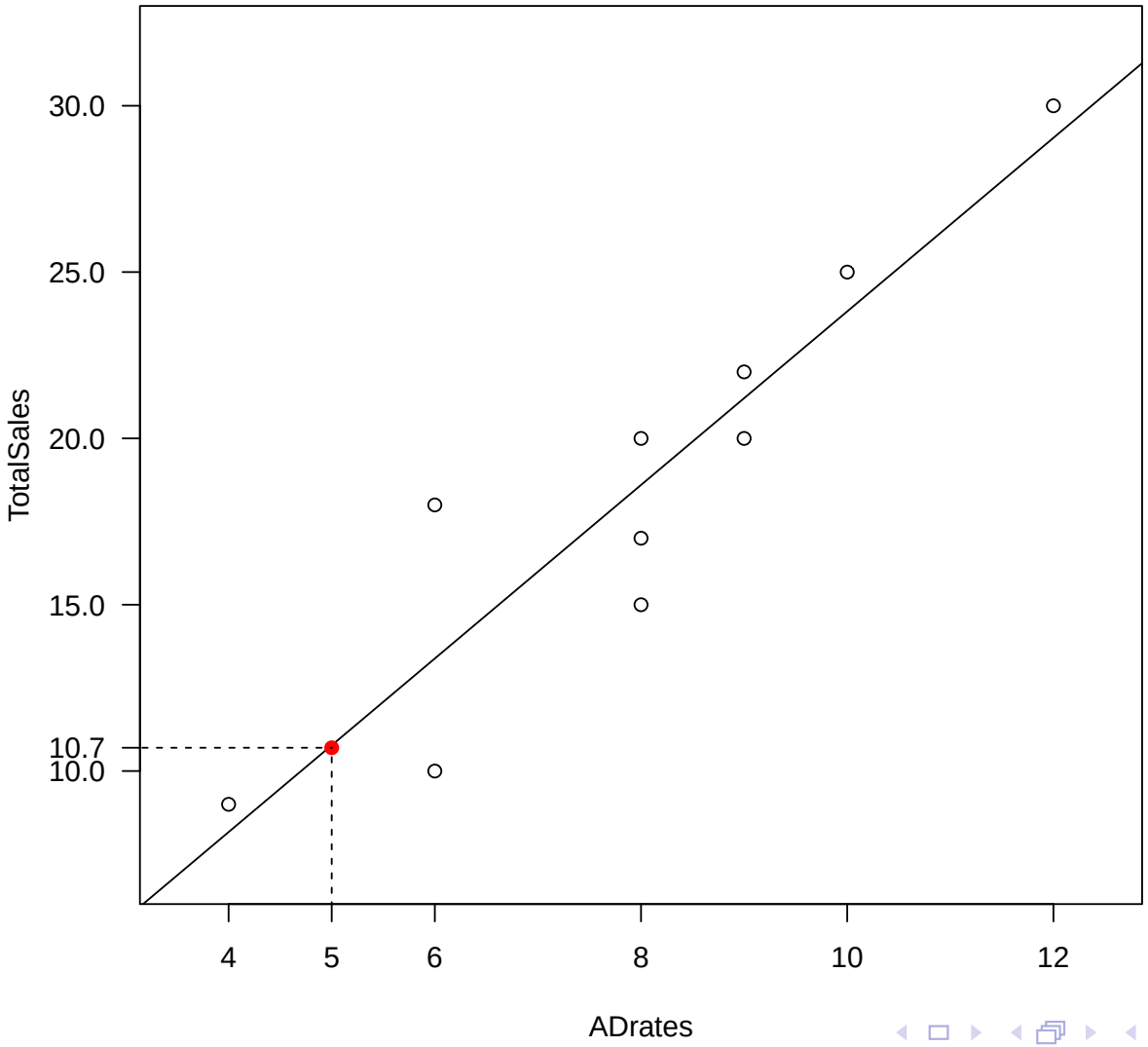
"응 광고료 50만원 쓰면 총 판매액은 1200만원 이야~"

배우신 분(Data Scientist)의 예측

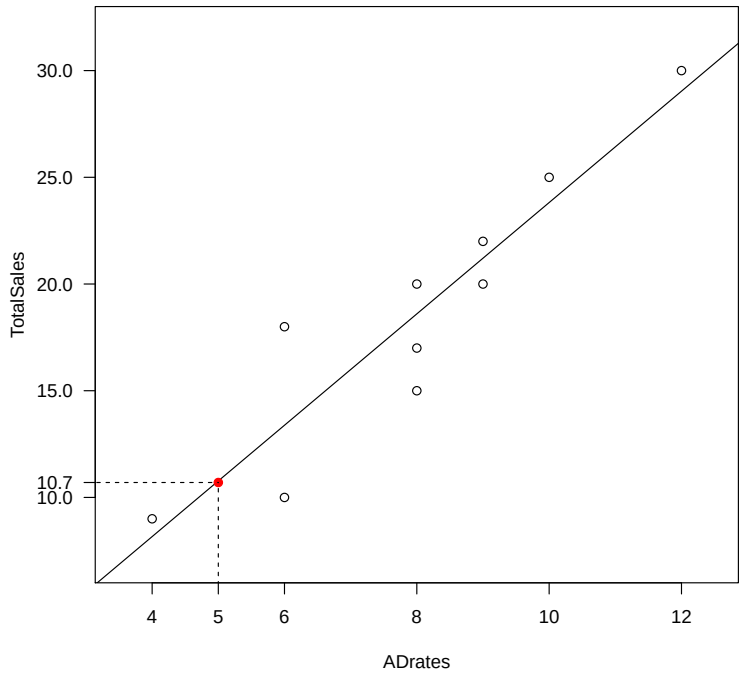


배우신 분(Data Scientist)의 예측

배우신 분(Data Scientist)의 예측

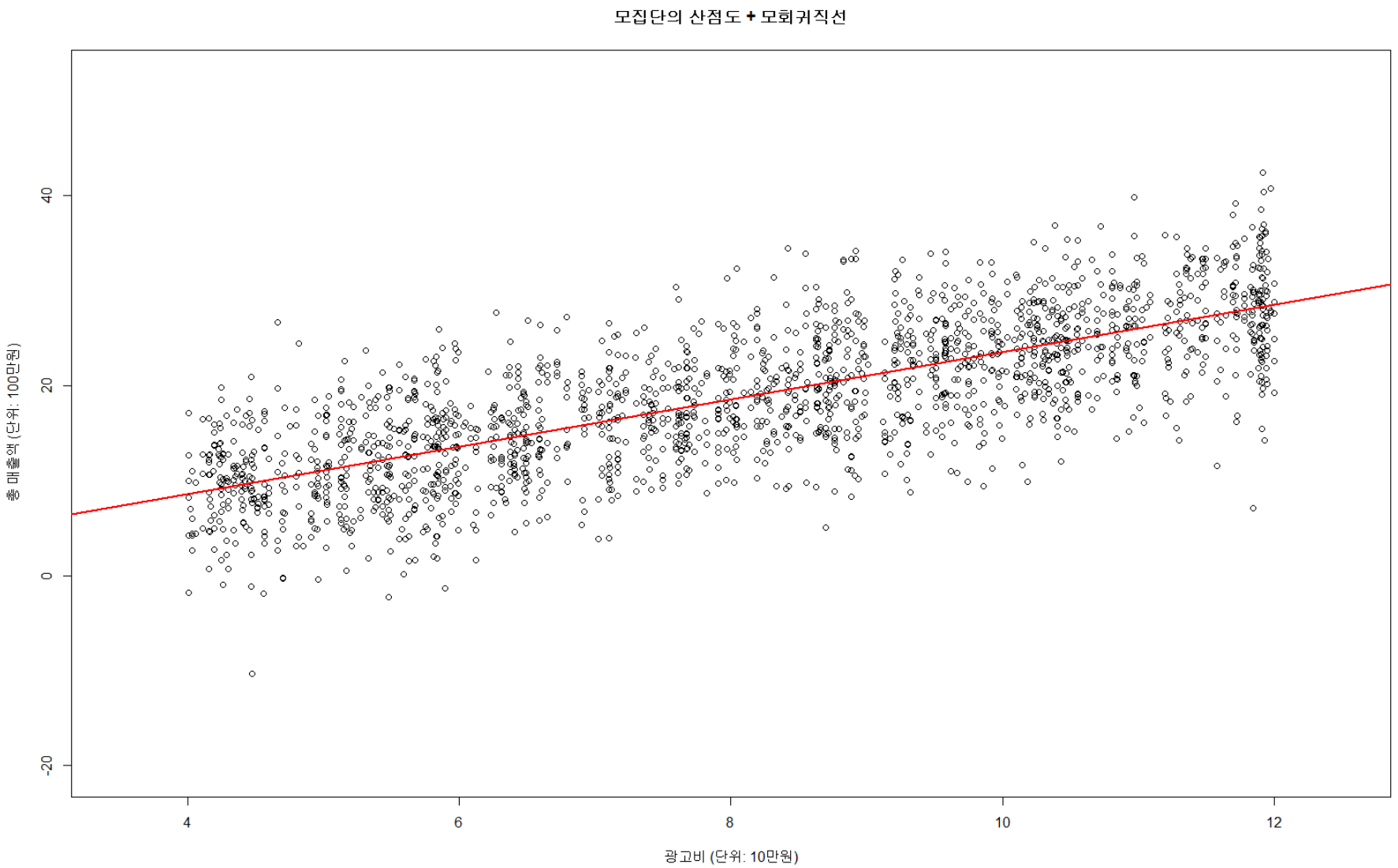


배우신 분(Data Scientist)의 예측

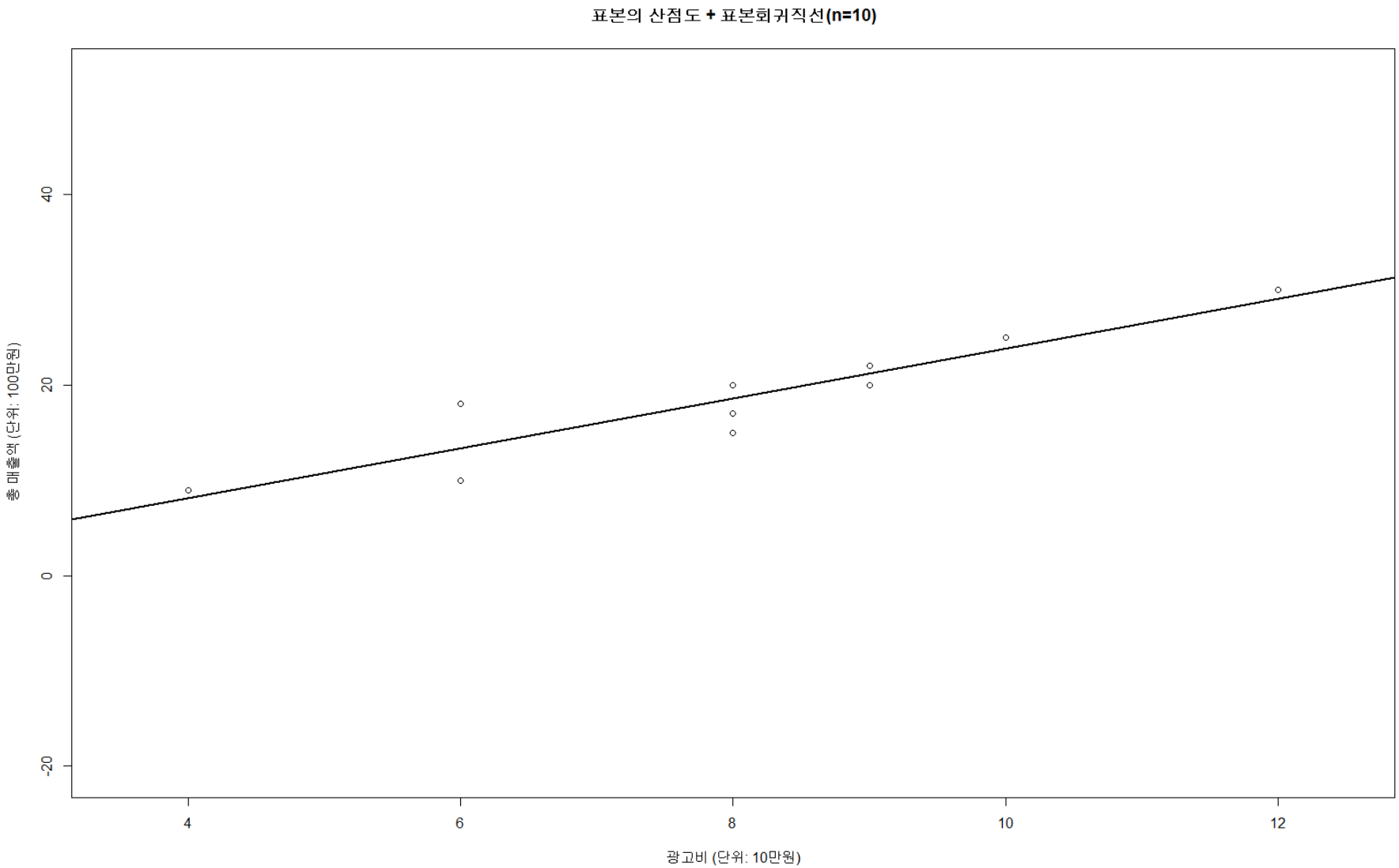


"광고료로 50만원을 사용하면, 총 판매액의 '**평균**'의 예측값은 1,007만원입니다."

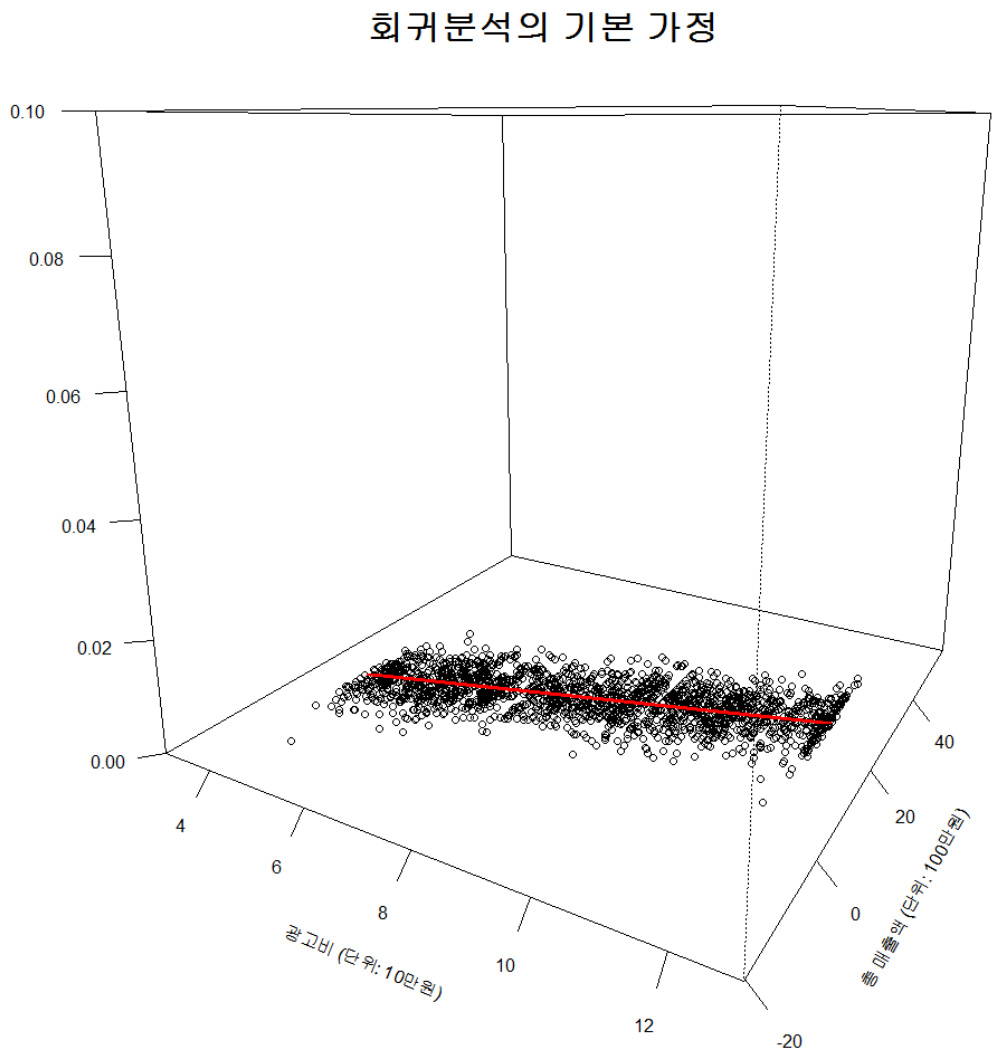
배우신 분(Data Scientist)의 예측



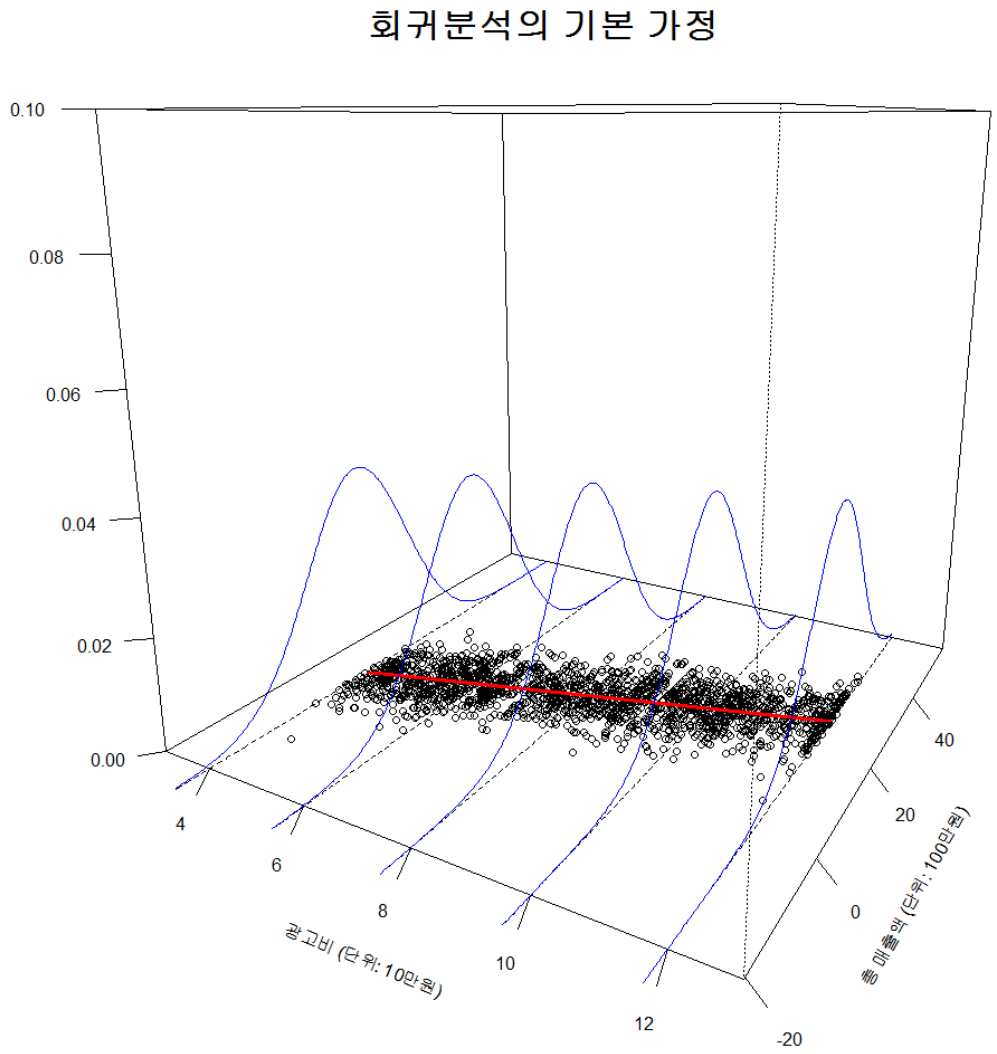
배우신 분(Data Scientist)의 예측



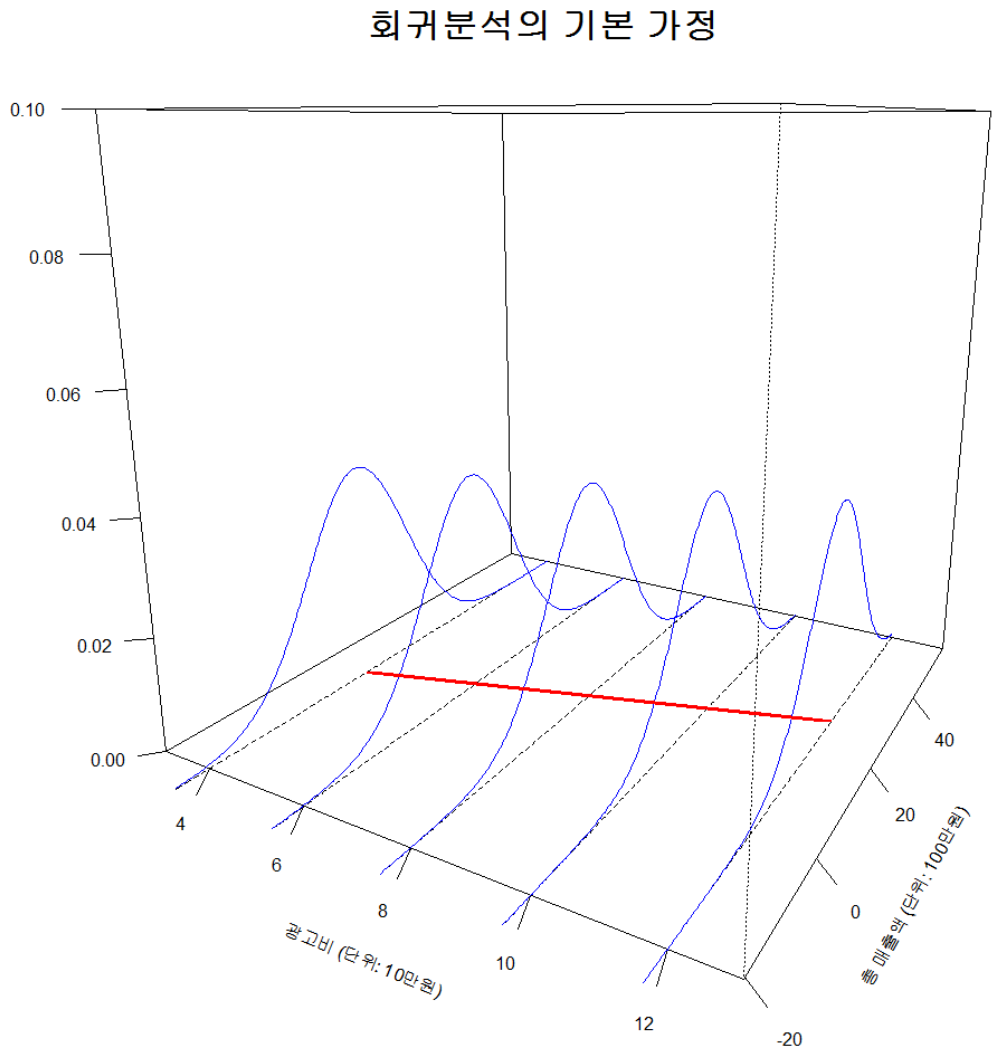
배우신 분(Data Scientist)의 예측



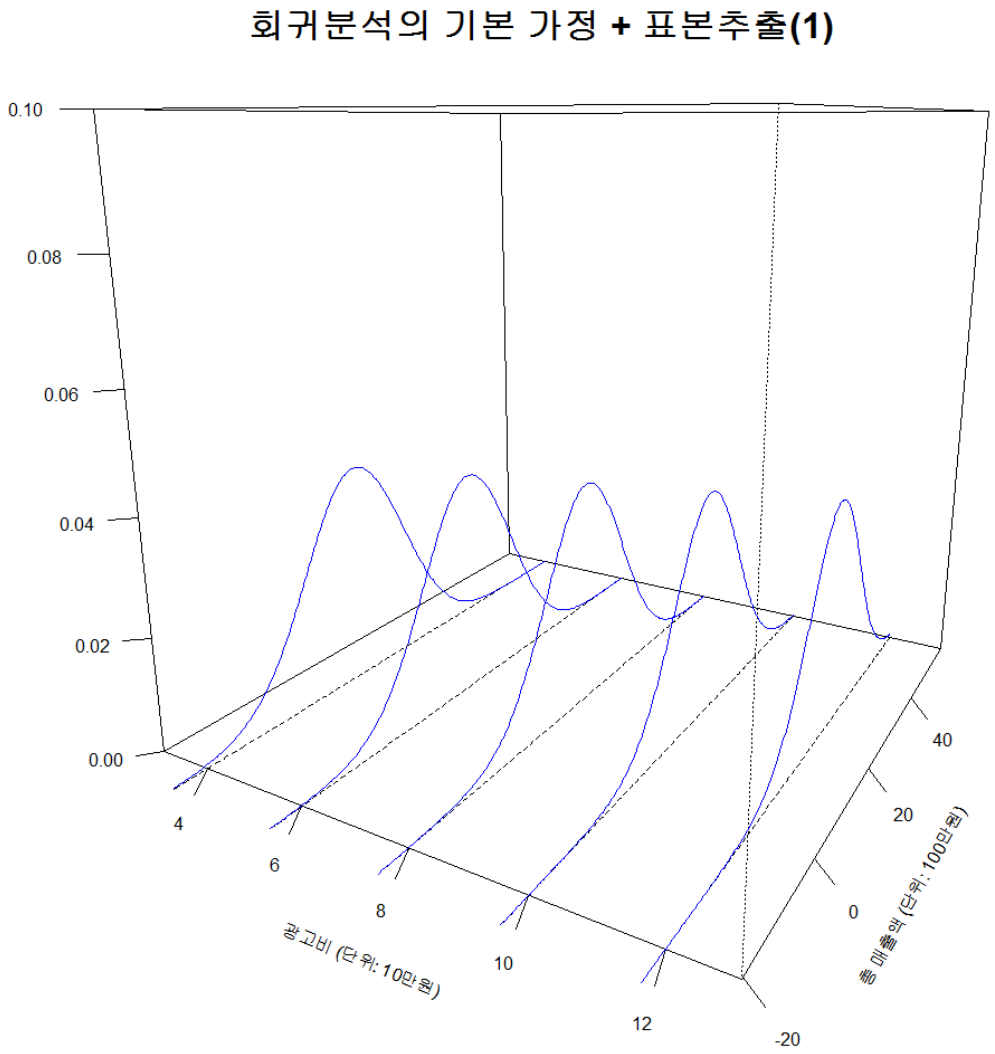
배우신 분(Data Scientist)의 예측



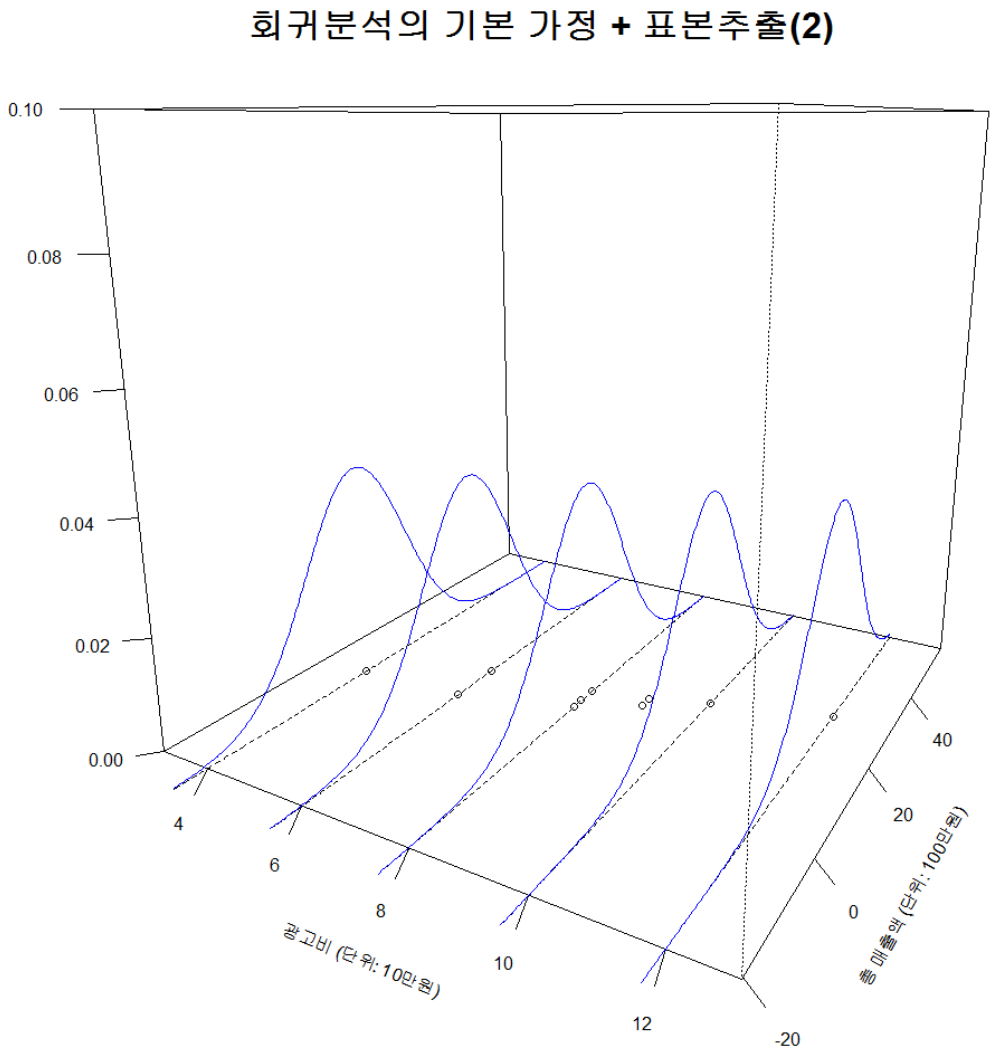
배우신 분(Data Scientist)의 예측



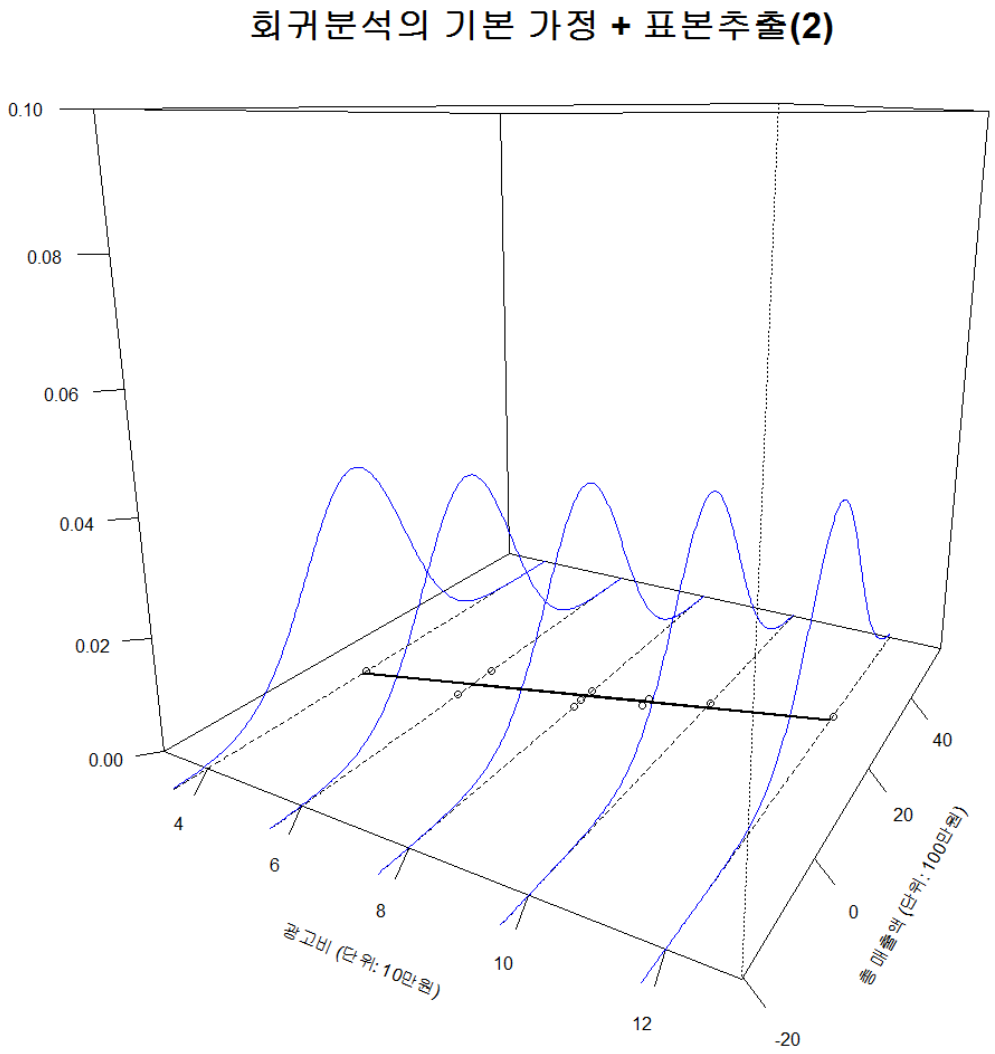
배우신 분(Data Scientist)의 예측



배우신 분(Data Scientist)의 예측



배우신 분(Data Scientist)의 예측



배우신 분(Data Scientist)의 예측

- 설명력?!

배우신 분(Data Scientist)의 예측

- 설명력?!
- 광고비와 총 매출액간의 선형회귀모형 적합 결과, 추정된 회귀식은

$$\text{총 매}\hat{\text{출}}\text{액} = -2.27 + 2.609\text{광고비}$$

이며, 이로부터 광고비가 한 단위(10만원) 오르면 총 매출액의 **'평균'**의 증가량은 2.609 (단위 : 100만원) 라는것을 알 수 있다.

배우신 분(Data Scientist)의 예측

- 설명력?!
- 광고비와 총 매출액간의 선형회귀모형 적합 결과, 추정된 회귀식은

$$\text{총 매}\hat{\text{출}}\text{액} = -2.27 + 2.609\text{광고비}$$

  이며, 이로부터 광고비가 한 단위(10만원) 오르면 총 매출액의 **'평균'**의 증가량은 2.609 (단위 : 100만원) 라는것을 알 수 있다.
- 오차의 정규분포 가정이 만족된다면 '**예측값의 표준오차**'를 구할 수 있는데, 이것이 바로 '**내가 예측한 값이 틀렸을 가능성**'까지 제시해주는 것이다.

배우신 분(Data Scientist)의 예측

- 광고비가 50만원 일 때, 예측값의 표준오차 $\sqrt{Var(\hat{Y})}$는

$$\sqrt{Var(\hat{Y})} = \sqrt{MSE\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{(xx)}}\right]}$$

$$= \sqrt{6.92\left[\frac{1}{10} + \frac{(5-8)^2}{46}\right]} = 1.76 \text{ (단위 : 100만원)}$$

이다. 이 값의 해석을 배우신 분의 최종 결론으로 알아보자.

배우신 분(Data Scientist)의 결론

- 광고비가 50만원 일 때 '**추정된 총 매출액의 평균**'은 1,007만원 이고, 이 값은 광고비가 50만원일 때 '**실제 총 매출액의 평균**'과 약 176만원 정도 다르다.

배우신 분(Data Scientist)의 결론

- 광고비가 50만원 일 때 '**추정된 총 매출액의 평균**'은 1,007만원 이고, 이 값은 광고비가 50만원일 때 '**실제 총 매출액의 평균**'과 약 176만원 정도 다르다.
- 또한 예측값의 95% 신뢰구간은

$$(747\text{만원},\ 1407\text{만원})$$

이다.