

4차 산업혁명 M·BA

비즈니스 애널리틱스.

회귀분석을 활용한 비즈니스 의사결정



LEARNING

회귀분석 개념과 단순회귀분석 예시

회귀분석이란?

회귀분석(Regression Model)이란

회귀분석 (Regression Model)

- 모든 사람, 데이터는 평균으로 접근하려는 복귀 성향이 있음을 표현
- 변수들 간의 관계를 선형(linear)으로 표현

회귀분석이란?

회귀분석(Regression Model)이란

Classification

분류하여
답을 찾음

회귀분석

종속변수의
연속형

예

골프 스코어, 학적, 주가

- * 연속변수 : 사람, 대상을 또는 사건을 속성의 크기나 양에 따라 분류할 수 있는 것을 의미

회귀분석이란?

회귀분석(Regression Model)이란

종속변수, 응답변수

예측되는 변수

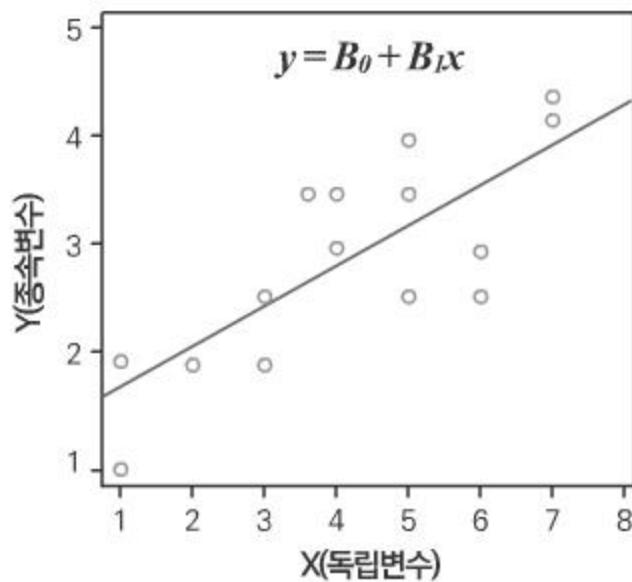
독립변수

예측에 활용되는 변수

회귀분석이란?

회귀분석(Regression Model)이란

인과관계 예측



X가 Y에 영향을 주는지 측정

X와 Y, 두 변수 간의 관계를
직선(Linear)형태로 표현

- 데이터 포인트 점을
가장 잘 나타내는
직선 긋기

회귀분석 모형

회귀선 추정과 추정 계수

회귀분석 모형

회귀분석 추정

- 표본의 관측개체에
가장 적합한 회귀선을
삿갓 모양(^\wedge)으로 표현

$$\hat{Y} = \hat{B}_0 + \hat{B}_1 X$$

추정된 절편 계수

추정된 기울기 계수

회귀선 추정과 추정 계수

$$Y = \boxed{B_0} + B_1 X$$

절편

- X(독립변수)의 값이 0일 때 Y(종속변수)의 값이 얼마인지, 얼마만큼 영향을 주는지 알려줌

회귀선 추정과 추정 계수

$$Y = B_0 + B_1 X$$

기울기

- 기울기는 X 가 한 단위 변할 때 Y 는 얼마나 변하는지 보여줌

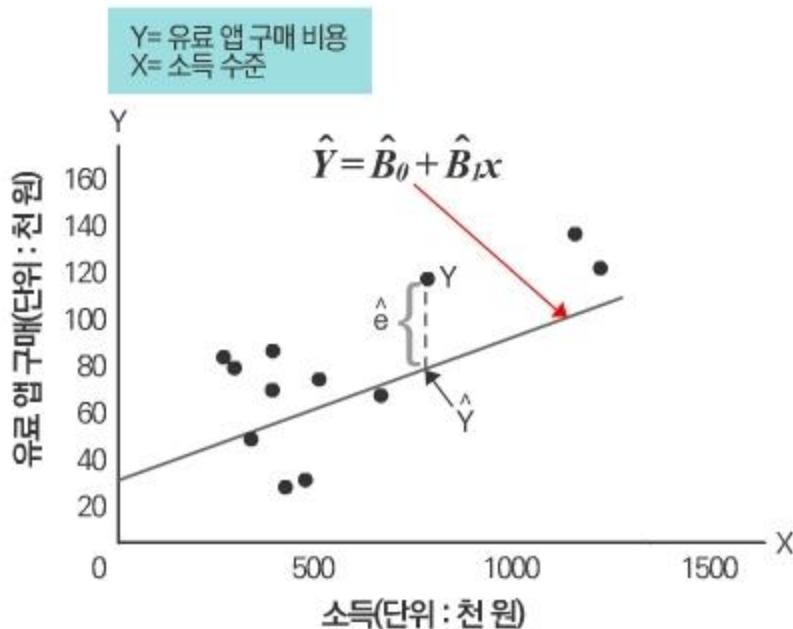
회귀선 추정과 추정 계수

$$Y = B_0 + 0.25X$$

- X가 1만큼 증가하면 Y는 0.25만큼 증가

회귀선 추정과 추정 계수

소득에 따라서 휴대폰 유료앱을 얼마만큼 구입하는가?



$$Y = B_0 + B_1 X$$

기울기

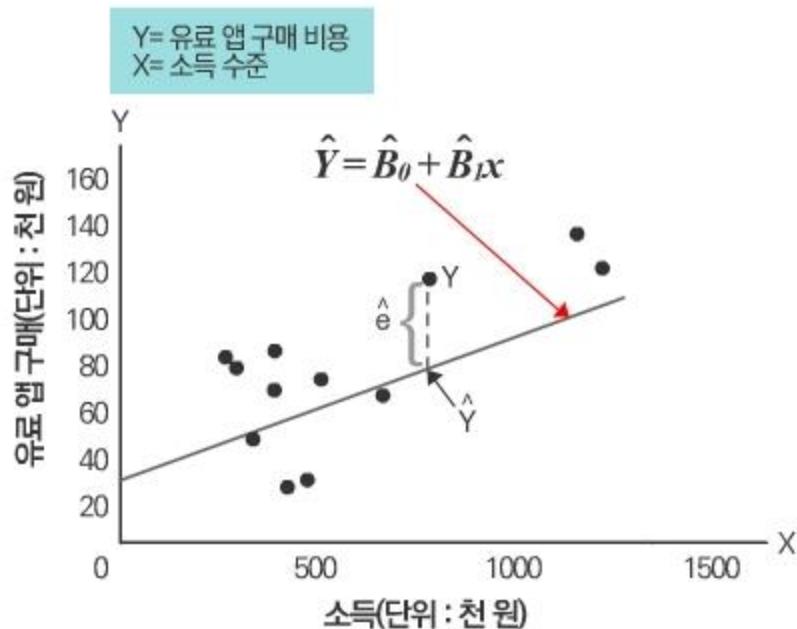
1만 원 증가

2500원 증가

* B_1 (기울기)은 0.25로 가정

회귀선 추정과 추정 계수

소득에 따라서 휴대폰 유료앱을 얼마만큼 구입하는가?



$$Y = B_0 + B_1 X$$

절편(X가 0일 때 Y값)

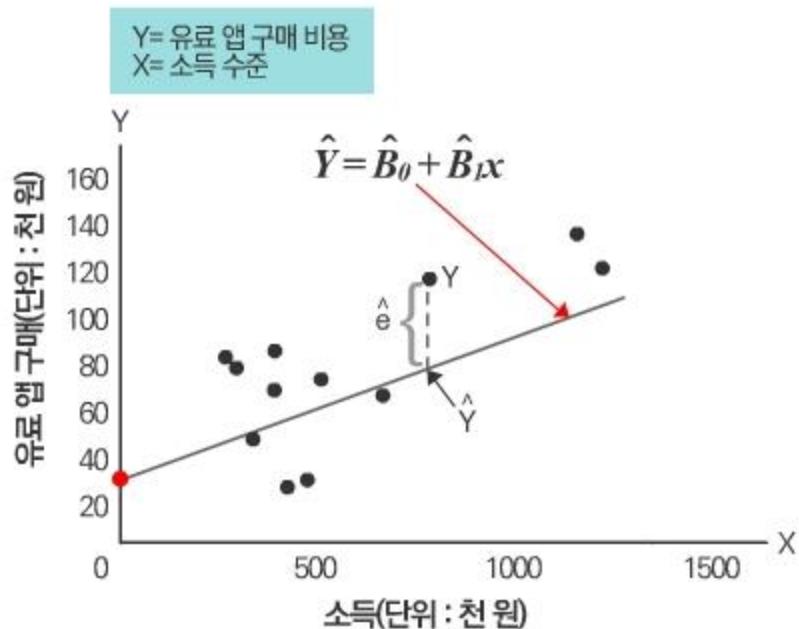
$$Y=0, B_0=Y$$

$$Y = B_0$$

어떤 특정 달에 소득이 없어도 유료 앱을 구매하는 값

회귀선 추정과 추정 계수

소득에 따라서 휴대폰 유료앱을 얼마만큼 구입하는가?



$$Y = B_0 + B_1 X$$

절편(X가 0일 때 Y값)

$$B_0 = 30$$

$$Y = B_0$$

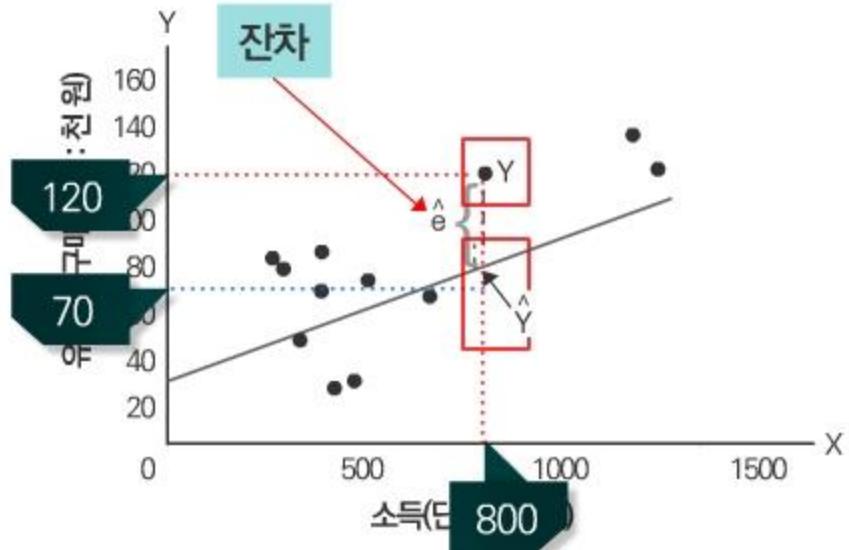
특정 달 소득 없이 유료 앱 구매 지출
3만 원

회귀선 추정과 추정 계수

잔차 (Residual)

$$\hat{e} = Y - \hat{Y}$$

- 모델을 통해서 예측하고, 설명하지 못하는 부분



모든 관측점들의 \hat{e}^2 의 합을 최소화하는 회귀선이 가장 적합한 직선이 됨

▶ 회귀분석 모형

회귀선 추정과 추정 계수

통상최소제곱법 (Ordinary least squares : OLS)

\hat{e}^2 의 합을 최소화하는 추정된 계수 \hat{B}_1 와 \hat{B}_0 의 값을 추정하는 방법

다중회귀분석 사례와 분석

다중회귀모델 추정

유료 앱 구매 다중회귀모델 회귀분석 결과

(단위 : 천 원)

변수	계수
절편	89.09***
INCOME	0.064**
PRICE	-3.18***
RAINFALL	6.05***

다중회귀분석 사례와 분석

다중회귀모델 추정

유료 앱 구매 다중회귀모델 회귀분석 결과

모든 독립변수의 값이 0일 때,
앱 구매 지출액을 나타냄

(단위 : 천 원)

계수

절편 \hat{B}_0

89.09***

- X 가 0일 때 앱을 구매하는 경우는 거의 없으므로 B_0 의 의미가 없어짐(변수가 여러 개) ***
- 앱 구매 평균 금액이 0일 수 없음 ***

다중회귀분석의 X 절편은 큰 의미가 없음

다중회귀분석 사례와 분석

다중회귀모델 추정

유료 앱 구매 다중회귀모델 회귀분석 결과

(단위 : 천 원)

변수	계수
평균 가격	1 증가
PRICE \hat{B}_2	지출 금액 3.18 감소 -3.18***

\hat{B}_2 의 값이 음의 부호임을 봤을 때, 앱 평균가격과 지출금액은 역의 관계를 가지고 있음을 알 수 있음

다중회귀분석 사례와 분석

앱 구매 모델의 종속변수 값 추정

유료 앱 구매 다중회귀모델 회귀분석 결과

(단위 : 천 원)

변수	계수
절편	89.09***
INCOME	0.064**
PRICE	-3.18***
RAINFALL	6.05***

$$\hat{APP EXP} = 89.09 + 0.064 * INCOME - 3.18 * PRICE + 6.05 * RAINFALL$$

다중회귀분석 사례와 분석

앱 구매 모델의 종속변수 값 추정

4월 관측개체 예상

(단위 : 천 원)

변수	추정값(X)
절편	89.09
INCOME	504
PRICE	18.99
RAINFALL	1.9

Y값 앱
구매 지출액

$$\hat{APP EXP} = 89.09 + 0.064 * 504 - 3.18 * 18.99 + 6.05 * 1.9 = 72.45$$

다중회귀분석 사례와 분석

추정된 종속변수의 값과 오차항

4월 실제
앱 구매 지출

73.040

4월 추정된
앱 구매 지출

72.450

$$\hat{e} = Y - \hat{Y}$$

$$\begin{aligned} &= 73.040 - 72.450 \\ &= 0.59 \end{aligned}$$

다중회귀분석 사례와 분석

다중회귀모델 추정

단순회귀분석(독립변수 : Income)

4월 실제
앱 구매 지출

73.4

4월 추정된
앱 구매 지출

71.34

$$e = Y - \hat{Y}$$

$$\begin{aligned} &= 73.4 - 71.34 \\ &= 1.70 \end{aligned}$$

4월 실제
앱 구매 지출

73.040

4월 추정된
앱 구매 지출

72.450

$$\begin{aligned} &\text{잔차} \\ &\hat{e} = Y - \hat{Y} \end{aligned}$$

$$\begin{aligned} &= 73.040 - 72.450 \\ &= 0.59 \end{aligned}$$

다중회귀분석

앱 구매 지출을
더 잘 설명

잔차(실제 값과 예측 값의 차이)가 적을 수록 더 좋은 모델임

회귀분석 시 여러 개의 독립변수를 활용하면 더 정확한 모델을 도출할 수 있음

잔차를 줄이는 방법

「좋은 모델을 만들기 위해
독립변수를 어떻게 활용할까?」



기존 자료에서 예측 시 필요한 독립변수 정보
탐색



Domain Knowledge 기반 딥러닝,
머신러닝을 통해 새로운 변수 생성

회귀계수 유의성 검정

‘계수에 대한 유의미성 검증이 필요’

유의함

유의하지 않음

- 독립변수가 종속변수에 영향을 줌
- 독립변수에 따라 종속변수가 어떻게 변하는지 알려줌
- 독립변수가 종속변수에 영향력이 없음

기울기 계수에 대한 유의성 검정

$$H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$$

(B_n 은 표본, β_n 은 추정 값)

$B_1 = 0$ (X 가 한 단위 증가해도
값에 상관없이 Y 값은 고정)

$$B_1 \neq 0$$

영향력 有(유의함)

영향력 無(유의하지 않음)

$$P < 0.05$$

$$P > 0.05$$

계산된 P값을 통해 계수의 유의성 여부를 판단할 수 있음

다중회귀분석 사례와 분석

다중회귀 분석 결과 예시

```
Console <--> 
> lm.res<-lm(formula=birth~w.act+edu.spd+cru.div+rGDPgw,data=Birth.df)
> summary(lm.res)

Call:
lm(formula = birth ~ w.act + edu.spd + cru.div + rGDPgw, data = Birth.df)

Residuals:
    Min      1Q  Median      3Q     Max  
-1.15217 -0.22191  0.07337  0.30661  0.82869

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 10.210782   1.746329   5.847 1.11e-06 ***
W.act        -0.206442   0.042851  -4.818 2.62e-05 ***
edu.spd       0.227689   0.063946   3.561  0.00106 ***
cru.div      -0.581267   0.199087  -2.920  0.00601 ***
rGDPgw      -0.006666   0.021069  -0.316  0.75354  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.456 on 36 degrees of freedom
Multiple R-squared:  0.7899, Adjusted R-squared:  0.7666 
F-statistic: 33.85 on 4 and 36 DF,  p-value: 9.644e-12
```

* 표시가 있으면
유의한 변수

조정결정계수

결정계수와 적합도

R^2
(R Square)

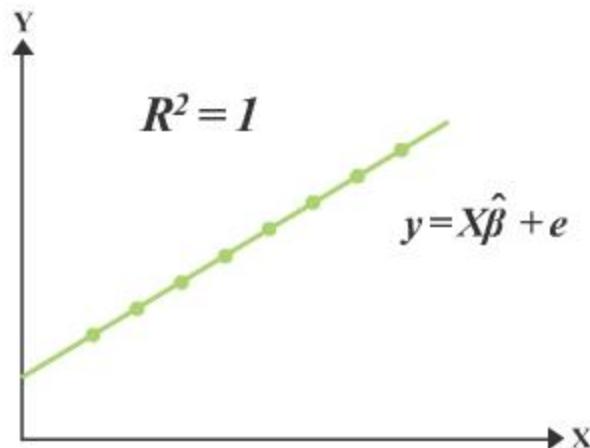
- 회귀모델 전체의 적합도 판단
- 표본회귀선이 표본자료를 얼마나 잘 반영하는지 판단할 도구로 활용
(표본회귀선의 설명력 측정)

$$0 \leq R^2 \leq 1$$

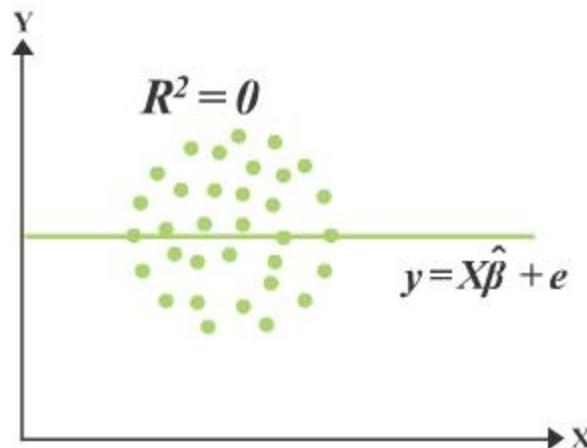
- 1과 가까울수록 좋은 모델

독립변수를 많이 넣을때 R^2 값이 상승하는 현상을
조절하기 위해 조정결정계수를 도입함

결정계수와 적합도



종속변수와 설명변수간에
선형관계가 존재하여
표본회귀선이 표본자료를
가장 잘 설명하고 있음



종속변수와 설명변수간에
선형관계가 없어
표본회귀선이 표본자료를
설명하지 못함

LEARNING

회귀분석 모델 측정



더미변수(dummy variable)

* 더미변수 : 측정의 편의상 사용하는 특정 변수

- 질적 효과를 고려할 수 있는 독립변수
- 질, 속성의 유무를 나타내는데 사용
- 변수 : 1,0

$$y_i = a_1 + \beta x_i + \varepsilon_i$$

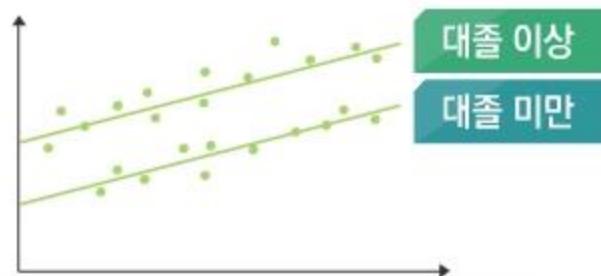
대출 이상 사원의 경우

$$= a_2 + \beta x_i + \varepsilon_i$$

대출 미만 사원의 경우

더미변수(dummy variable) 사례 1

학력 차이로 초임이 다른 경우
임금과 근무 연수의 관계

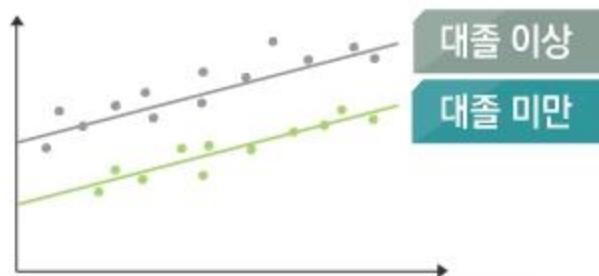


$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \varepsilon_i$$

1 : 대졸 이상, 0 : 대졸 미만

더미변수(dummy variable) 사례 1

학력 차이로 초임이 다른 경우
임금과 근무 연수의 관계



$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \varepsilon_i$$

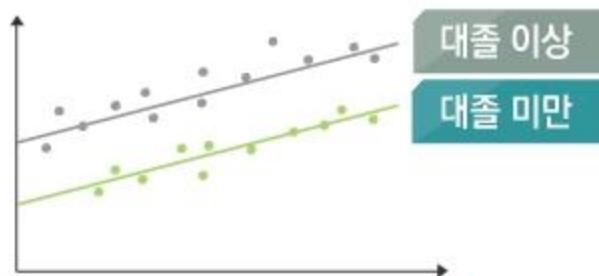
0

근무 년 수

근무 년 수가 증가할수록 임금은 상승함

더미변수(dummy variable) 사례 1

학력 차이로 초임이 다른 경우
임금과 근무 연수의 관계

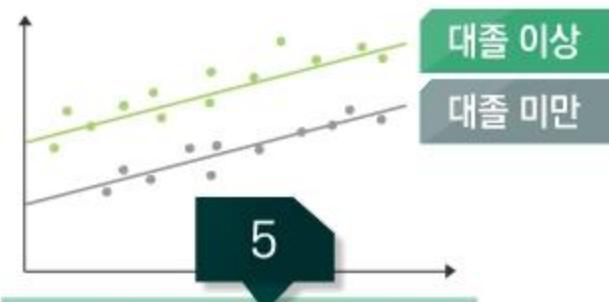


$$y_i = \beta_0 + \beta_2 X_i = 10$$

10

더미변수(dummy variable) 사례 1

학력 차이로 초임이 다른 경우
임금과 근무 연수의 관계



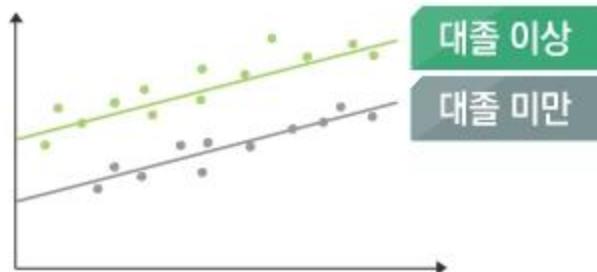
$$y_i = \beta_0 + \beta_1 D_i = 15$$

10

1

더미변수(dummy variable) 사례 1

학력 차이로 초임이 다른 경우
임금과 근무 연수의 관계



대졸 미만 사원의 절편

$$y_i = B_0 + B_2 X_i = 10$$

대졸 이상 사원의 절편

$$y_i = B_0 + B_2 X_i = 15$$

근무 년 수가 증가할수록 두 그룹 간의 연봉 차이가 동일하게 나타남

더미변수(dummy variable) 사례 2 – 기울기 조정

- 학력에 따라 임금 상승률 차이를 기울기로 표현

$$y_i = r_0 + r_1 D_{i1} + r_2 x_i + r_3 D_{i2} x_i + \varepsilon_i$$

0

0

대출 미만

$$y_i = r_0 + r_2 = 0.25$$

10

0.25

더미변수(dummy variable) 사례 2 – 기울기 조정

$$y_i = r_0 + r_1 D_{i1} + r_2 x_i + r_3 D_{i2} x_i + \varepsilon_i$$

1

0

대출 이상

$$y_i = r_0 + r_2 = 15$$

10

5

초임 임금차이 = 기울임 값

더미변수(dummy variable) 사례 2 – 기울기 조정

$$y_i = r_0 + r_1 D_{i1} + r_2 x_i + r_3 D_{i2} x_i + \varepsilon_i$$

1

0

대출 이상

$$y_i = (r_2 X + r_3) * X = 0.35X$$

0.25

0.1

초임 임금차이 = 기울임 값

더미변수(dummy variable) 사례 2 – 기울기 조정

대졸 미만 사원의 기울기

$$y_i = r_0 + r_2 = 0.25$$

대졸 이상 사원의 기울기

$$y_i = (r_2X + r_3) * X = 0.35$$

더미 개수

변수 Level -1

(예 : 그룹3으로 나눌 경우 더미 변수는 2개로 만들기)

* 다중공선성 : 회귀 분석에서 사용된 모형의 일부 예측 변수가 다른 예측 변수와 상관 정도가 높아, 데이터 분석 시 부정적인 영향을 미치는 현상

더미변수 사용 시 완전한 다중공선성의 문제가 발생되지 않도록 주의해야 함

선형회귀(Linear Regression)를 이용한 곡선추정

다항모델 (Polynomial Model)

- 곡선을 추정할 수 있음
(제곱, 세제곱 등
어떤 지수 형태이든
독립변수로 사용 가능)

다항모델의 예

$$Y = B_0 + B_1 X_1 + B_2 (X_1)^2 + B_3 X_3 + e$$

$$Y = B_0 + B_1 X_1 + B_2 (X_1)^2 + B_3 (X_1)^3 + e$$

스포츠카 생산비용 예측

- 평균비용함수를 추정하기 위해 스포츠카 다행모델 설정

(단위 : 천 원)

종속변수 AVECOST

한 대당 평균비용

독립변수 CARS

주당 생산되는 자동차 대수

관측개체	CARS	AVECOST
1	50	42,250
2	100	40,825
...
20	1,000	74,586

스포츠카 생산비용 예측 - 제곱합 곡선 예측

자동차 공장에서 1대 만드는 가격과
100대 만드는 가격이 같을까?

스포츠카 다향모델 회귀분석 결과

변수	계수	표준 오차	t - 통계량	유의 확률
절편	51,671.10	2,870.231	18.002	0.001
CARS	-121.662	12.590	-9.664	0.001
(CARS) ²	0.142	0.0116	12.224	0.001

자동차 생산량 ↑

평균 생산비용가격 ↓

공장 용량 이상의 생산 ↑

평균 생산비용가격 ↑

회귀분석 모델 측정 방법

스포츠카 생산비용 예측 - 제곱합 곡선 예측

변수	계수	표준 오차	t - 통계량	유의 확률
절편	51,671.10	2,870.231	18.002	0.001
CARS	-121.662	12.590	-9.664	0.001
(CARS) ²	0.142	0.0116	12.224	0.001

