

HEARTCOUNT 월간 웨비나 시리즈

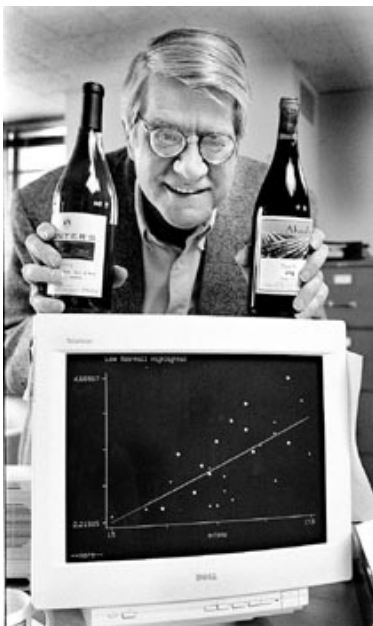
1. 하트카운트 시작하기 - 데이터셋 연동
2. 시계열 데이터 분석
3. 여러 차원을 한 화면에 시각화하기
4. 상관관계와 분포 - 개별 레코드 수준에서의 시각화 기법들
5. 드릴다운과 트리맵 - 전체를 구성하는 개별 요소들 확인하기
6. 분석하기 좋은 데이터셋을 구성하는 요소들
7. 특정 행동을 보인 집단의 특성을 이해하기
- 8. 회귀 분석을 통해 지표 차이의 요인 이해하기**
9. 상관관계로 인과성에 대해 이야기하는 법
10. LLM(거대언어모델)과 데이터 분석의 자동화
11. 데이터 보고서 잘 쓰는 법 (feat. 데이터 스토리텔링)
12. 월은 웨비나 쉽니다. 😊

양 승 준 / sidney.yang@idk2.co.kr

- 왜 고릿적 회귀분석인가?
- 모형에 대한 직관(Intuition) 키우기
- 적용, 해석할 때 고려할 점
- 간단한 시연

아주 단순하고 투명한 모형

단순하고 투명해서(White Box)
모형에 담긴 질서를 이해하고 언어로 옮기기 쉬움



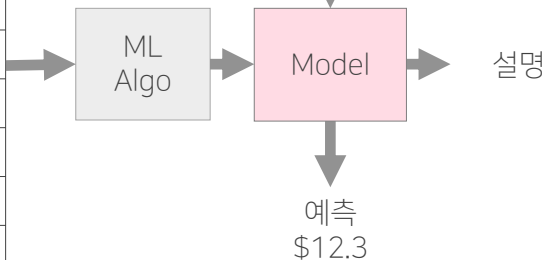
선형회귀 알고리즘
$$Y = a + bX_1 + cX_2 + dX_3$$

Ashenfelter's Wine Formula
$$\text{Price} = 12 + (0.1 \times \text{겨울강수량}) + (0.6 \times \text{평균온도}) - (0.4 \times \text{수확철강수량})$$

Training Data Set

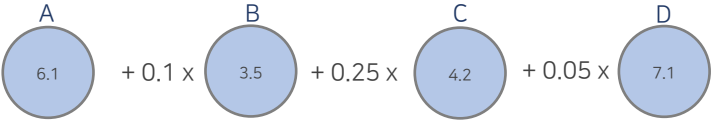
X ₁	X ₂	X ₃	Y
겨울강수량	수확철강수량	평균온도	와인가격
13	35	35	9.5
22	25	25	3.5
25	21	21	3.2
11	18	18	3.5
47	45	45	4.7
.	.	.	.

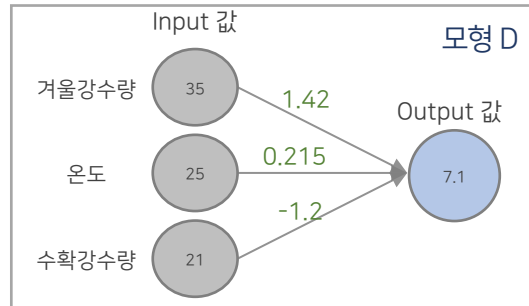
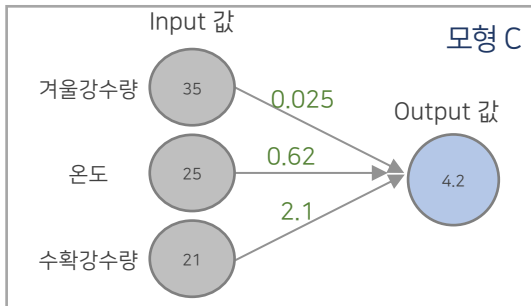
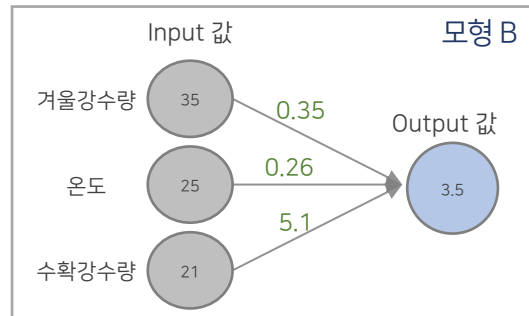
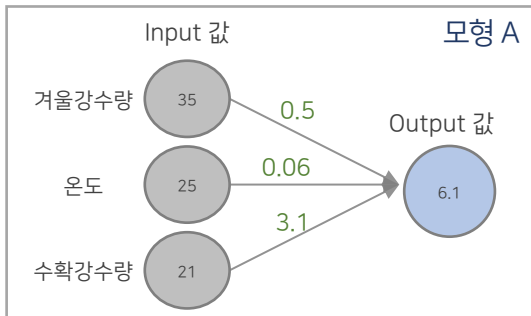
겨울강수량	수확철강수량	평균온도	와인가격
35	25	21	?



White Box vs. Black Box

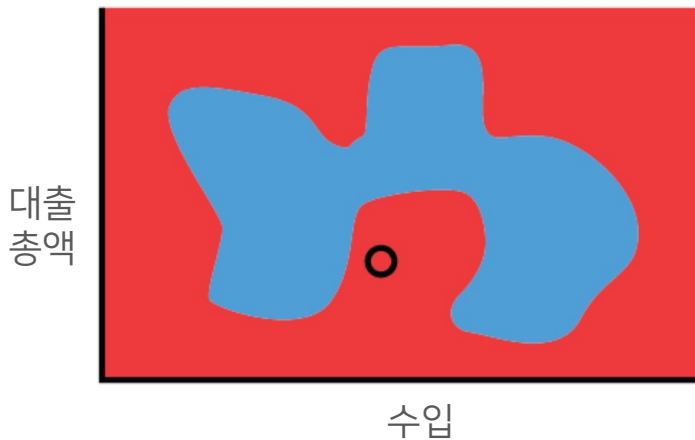
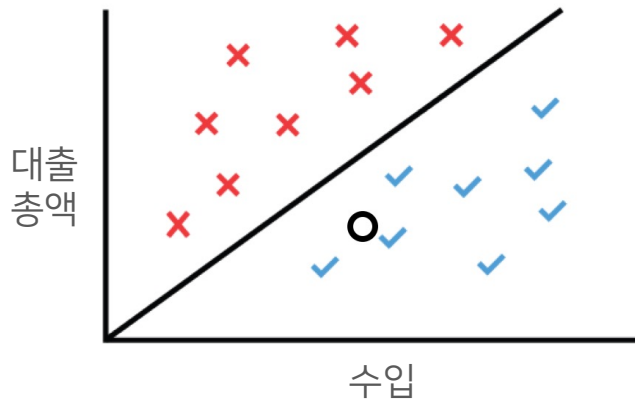
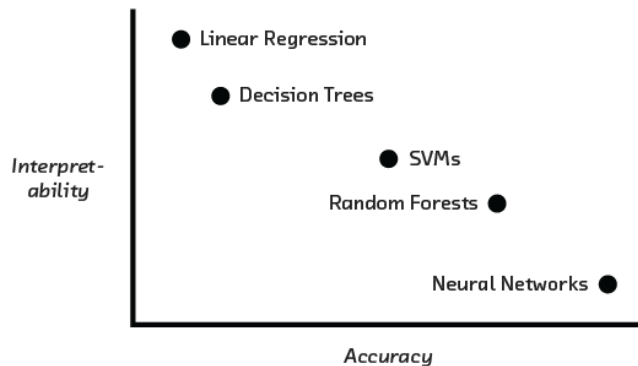
Black Box: 더 정확하지만 이해하기 어려움

$$\text{가격} = 15.4\$ = 0.6 \times \text{A} + 0.1 \times \text{B} + 0.25 \times \text{C} + 0.05 \times \text{D}$$




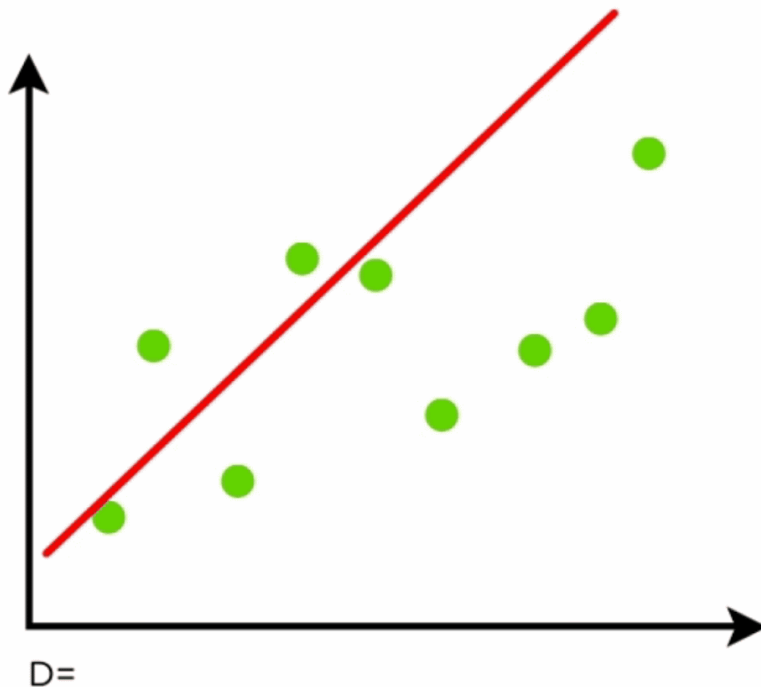
Model Accuracy vs. Interpretability

예측 vs. 설명
정확함 vs. 올바름



선형회귀분석 (Linear Regression Analysis)

Y값의 차이를 설명하기 위해 최선의 직선을 하나 긋는 일

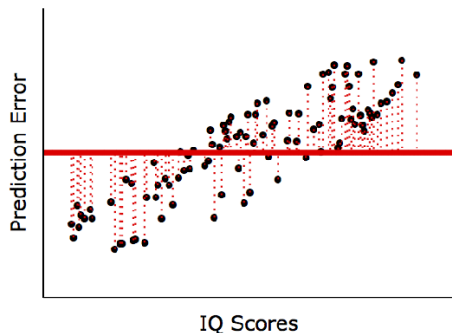
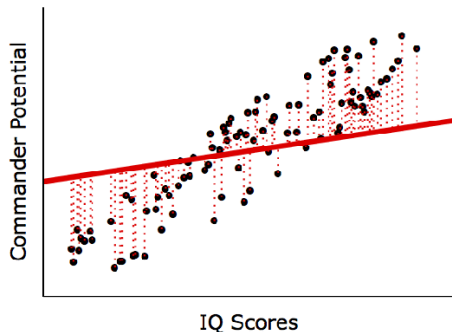


선형회귀분석 (Linear Regression Analysis)

Supervised
Machine-Learning

Regression Model: Y가 숫자형 변수(매출)인 경우

Classification Model: Y가 범주형 변수(성별)인 경우



Linear Regression

- 가장 오래되고, 널리 쓰이고, 이해하기 쉬운 지도학습 알고리즘
- 독립변수(X)로 숫자형 종속변수(Y)를 가장 잘 설명·예측(Best Fit)하는 선형 관계(Linear Relationship)를 찾는 방법 중 하나

계산방법 (Least Squares)

X와 Y 사이에 선형적 관계가 있다는 가정 하에
실제 Y값과 예측한 Y값 차이(Residual)를 최소화하는 방정식 계산

$$Y = b_0 + b_1X + \text{error}$$

- b_0 : Y축 절편(Intercept); 예측변수가 0일 때 기대 점수
- b_1 : 기울기, X가 한 단위 증가했을 때 Y의 평균적 변화값

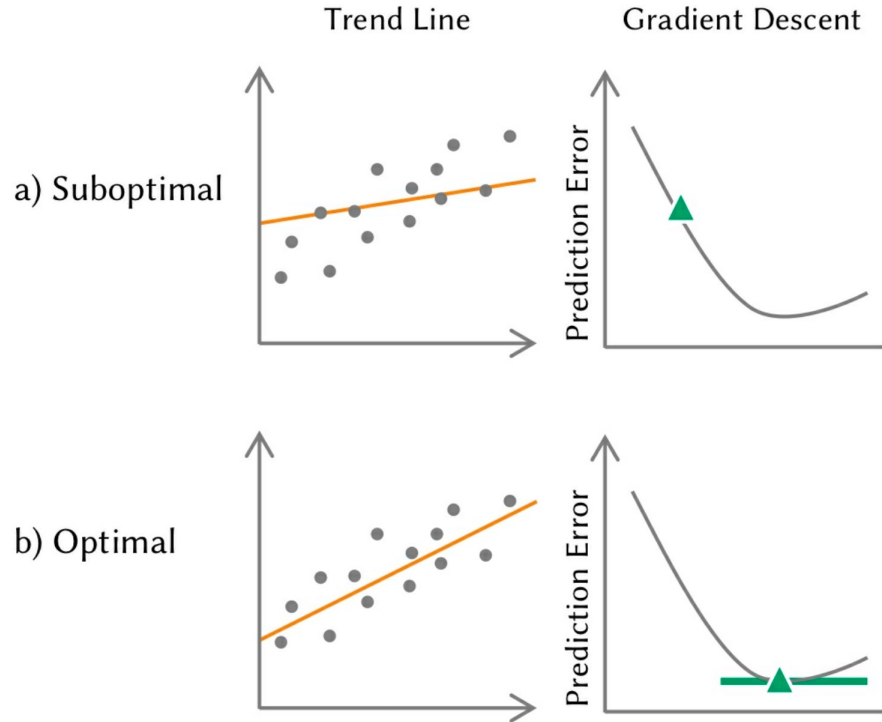
Least Square vs. Gradient Descent

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

$$\begin{aligned} RSS &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 \end{aligned}$$

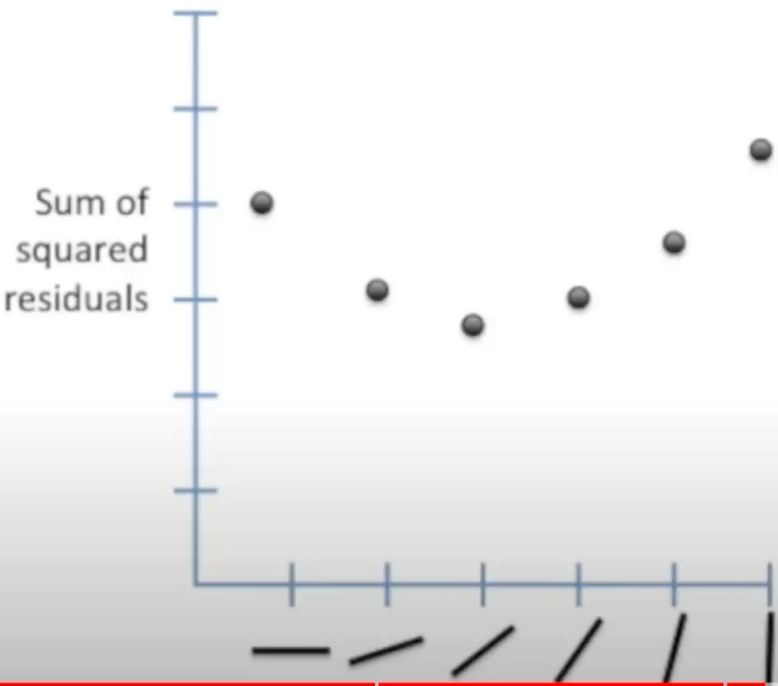
$$\hat{b}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$



Gradient Descent (경사하강법)

If we plotted the sum of squared residuals vs. each rotation, we'd get something like this...



6:21 / 9:21 • Least Squares > [Play] [CC] [Settings] [Fullscreen] [Close]

```
lm.fit = lm(sales ~ TV, data=ads)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = ads)
##
## 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 freed
om
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.60
99
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-1
6
```

RMSE/RSE

Q. 실제값과 예측값 사이에 얼마나 차이가 나는가?

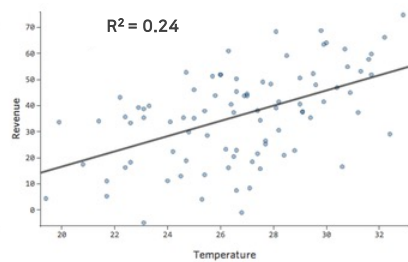
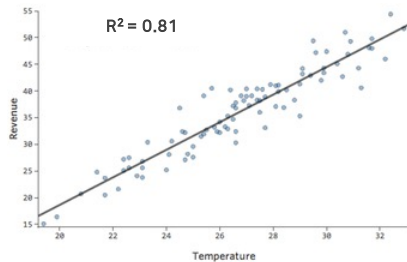
P-Value (Probability-Values)

Q. X와 Y 사이에 통계적으로 유의미한 관계가 있나?

- 관계의 세기(Size of an Effect)를 나타내는 것은 아님

R² (R-SQUARED; 결정계수)

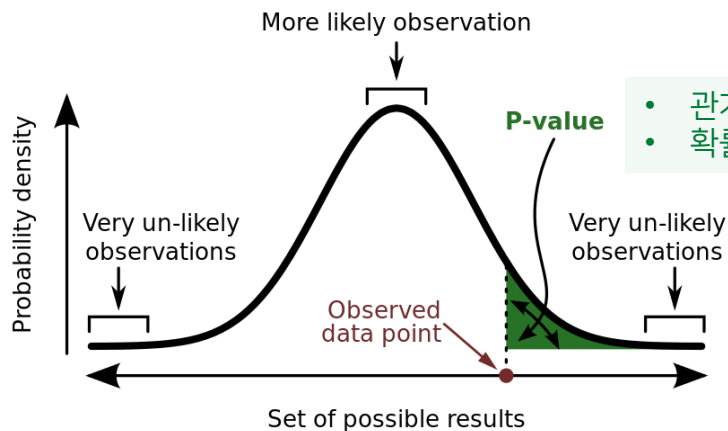
Q. X가 Y 변화를 얼마나 잘 설명하는가?



P-Value (Probability-Values)

P값: (선형적) 관계가 우연히 나왔을 확률
작은 P값: 데이터에서 발견한 관계가 우연이 아니다 (=통계적으로 유의미)

- 귀무 가설(H_0 ; Null Hypothesis): X와 Y 사이에 선형적 관계가 없다.
- $P = 0.05$: 관계가 없던 가정 하에 데이터에서 발견한 관계 혹은 그것보다 더 강한 관계가 관측될 확률 = 5%

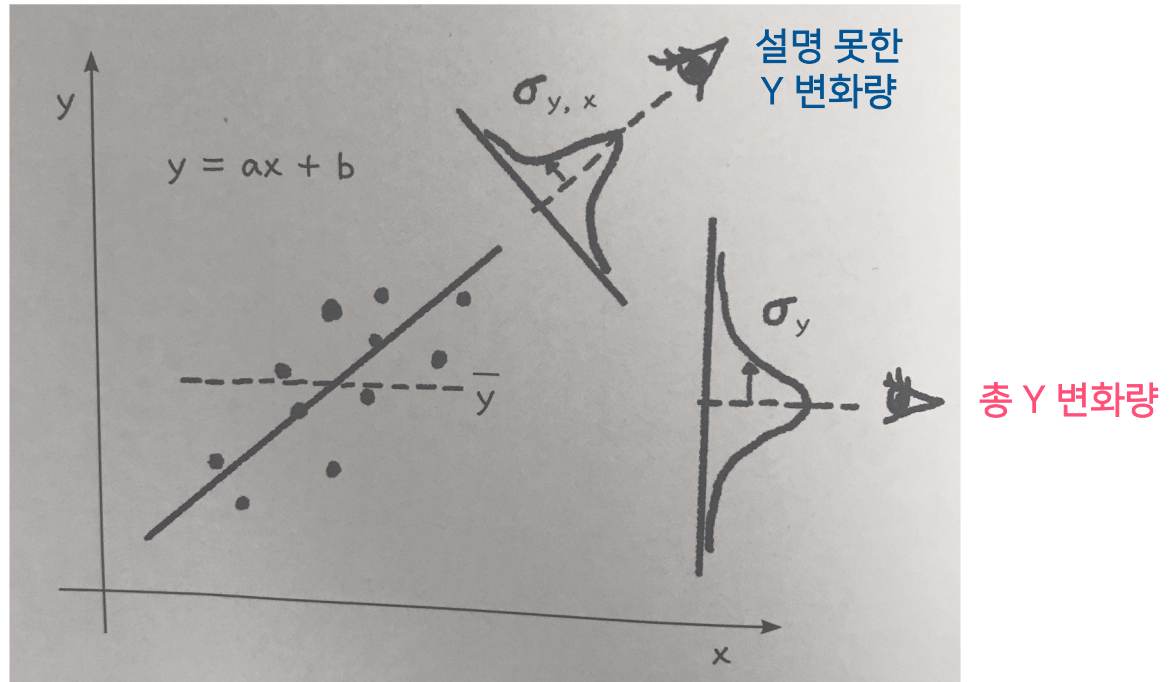


- 관계가 없을 때 해당 관계가 우연히 관찰될 확률
- 확률이 5%보다 작으면 관계가 있다고 결론

선형회귀분석: 결정계수(R²: R-SQUARED)

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

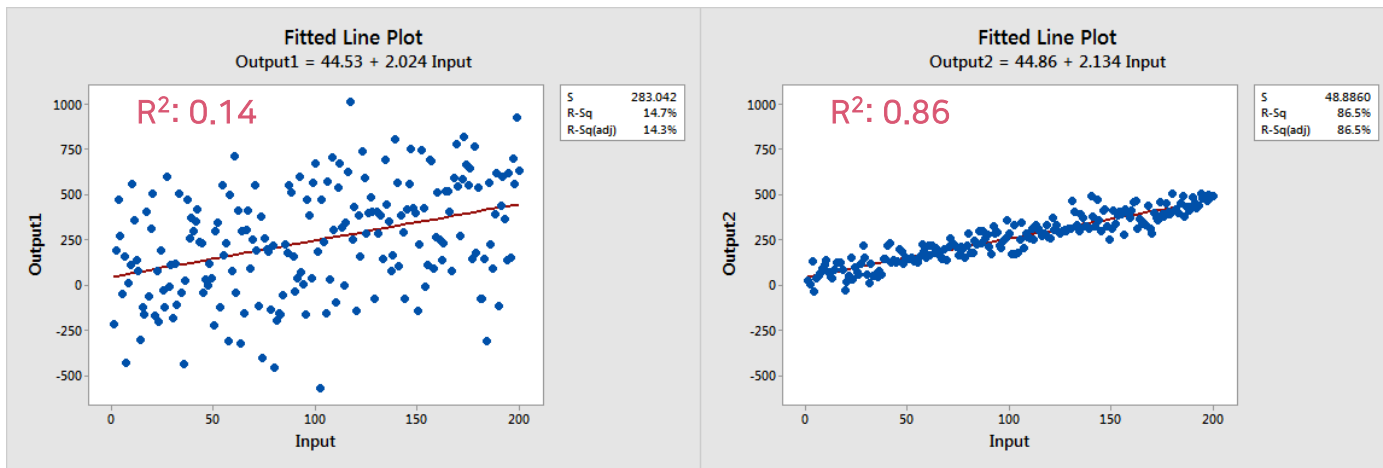
R² = 설명한 Y 변화량 / 총 Y 변화량
 = (총 Y 변화량 - 설명 못한 Y 변화량) / 총 Y 변화량
 = 1 - (설명 못한 Y 변화량 / 총 Y 변화량)



선형회귀분석: 결정계수(R²: R-SQUARED)

낮은 결정계수가 반드시 나쁜 (Inherently Bad) 것은 아님

- 동일한 회귀방정식: $Y = 44 + 2 * X$; $P < 0.001$
- 우측 모형이 좌측 모형보다 예측 정확도(R²)는 매우 높음
- 변수 간 경향성은 동일: X 1단위 증가 → Y 2단위 증가



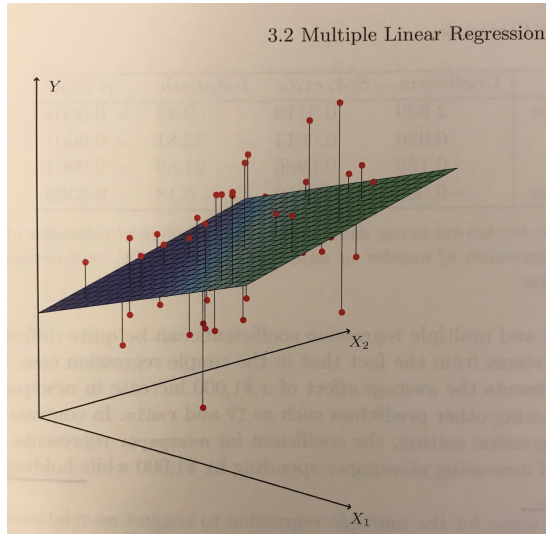
Y값을 정확히 예측하기 위해선 R² 값이 중요
하지만, 경향성 정보가 중요한 경우 R²가 낮다고 꼭 나쁜 모형은 아님

Multiple Linear Regression Analysis – Advertisement

변수 2개[TV, Radio]를 사용하여 Sales와의 관계를 설명·예측하는 회귀모형

$$Y = b_0 + b_1X_1 + b_2X_2$$

Sales = 2.9 + 0.045 x TV + 0.187 x Radio



SUMMARY OUTPUT				
<i>Regression Statistics</i>				
Multiple R		0.94720339		
R Square		0.897194261		
Adjusted R Square		0.896150548		
Standard Error		1.681360913		
Observations		200		
<i>ANOVA</i>				
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Regression	2	4860.2348	2430.1174	859.6177183
Residual	197	556.91398	2.8269745	
Total	199	5417.1488		
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	2.921099912	0.2944897	9.9191929	4.56556E-19
X Variable 1	0.045754815	0.0013904	32.908708	5.43698E-82
X Variable 2	0.187994227	0.00804	23.382446	9.77697E-59

선형회귀분석은 Causality 신경 안 씀

```
model = sm.OLS(train.y, train[['t']])
results = model.fit()
print(results.summary())
```

Test RMSE = 10

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared:                0.019
Model:                  OLS    Adj. R-squared:           0.019
Method:                 Least Squares    F-statistic:             2.766e+04
Date:                   Wed, 13 Feb 2019  Prob (F-statistic):      0.00
Time:                   07:25:25    Log-Likelihood:          -5.2235e+06
No. Observations:      1401801    AIC:                     1.045e+07
Df Residuals:          1401800    BIC:                     1.045e+07
Df Model:               1
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
t	0.9984	0.006	166.317	0.000	0.987	1.010

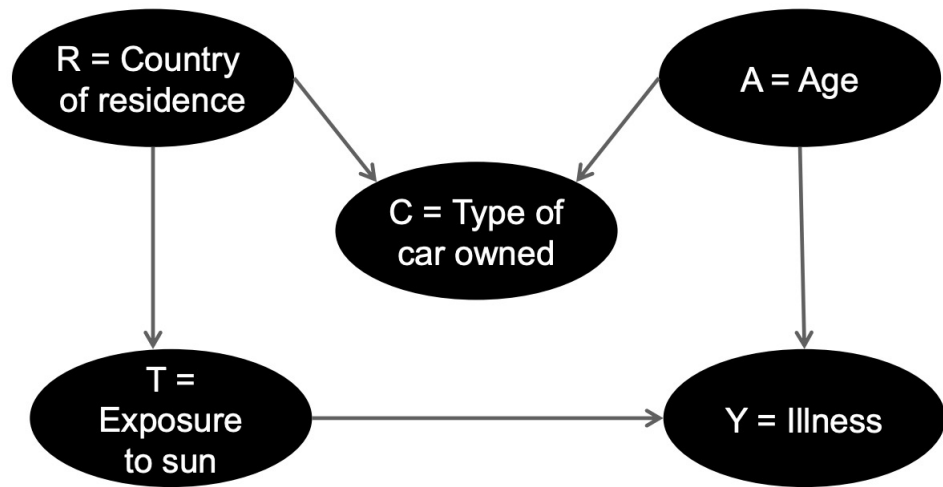
```
model = sm.OLS(train.y, train[['t', 'car']])
results = model.fit()
print(results.summary())
```

Test RMSE = 7.7

OLS Regression Results

```
=====
Dep. Variable:          y      R-squared:                0.408
Model:                  OLS    Adj. R-squared:           0.408
Method:                 Least Squares    F-statistic:             4.835e+05
Date:                   Wed, 13 Feb 2019  Prob (F-statistic):      0.00
Time:                   07:27:29    Log-Likelihood:          -4.8695e+06
No. Observations:      1401801    AIC:                     9.739e+06
Df Residuals:          1401799    BIC:                     9.739e+06
Df Model:               2
Covariance Type:       nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
t	-1.0036	0.005	-196.455	0.000	-1.014	-0.994
car	4.0024	0.004	959.740	0.000	3.994	4.011



$$R = \varepsilon_1$$

$$A = \varepsilon_2$$

$$C = R + A + \varepsilon_3$$

$$T = R + \varepsilon_4$$

$$Y = T + 10A + \varepsilon_5$$

$\varepsilon_1, \dots, \varepsilon_5$ independent normal(0,1)

Goal:
Find the effect of sun exposure on the illness

Binning: 선형회귀분석으로 비선형적 관계 찾기

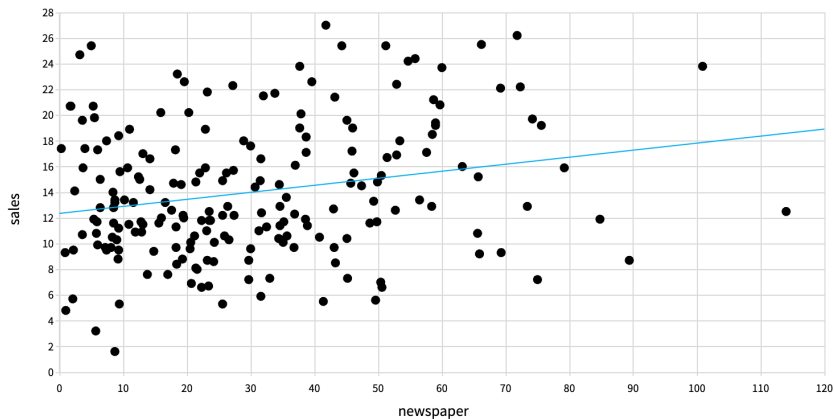
No.	variable name	R ² ⓘ	Adjusted R ² ⓘ	P-Value ⓘ	count	
1	TV	0.612	0.61	0.00000 (< 0.001 ***)	200	⌵
2	radio	0.332	0.329	0.00000 (< 0.001 ***)	200	⌵
3	newspaper	0.052	0.047	0.00115 (< 0.01 **)	200	⌵

Binning: 숫자형 변수를 범주형 변수로 변형

No.	variable name	R ² ⓘ	Adjusted R ² ⓘ	P-Value ⓘ	count	
1	TV	0.612	0.61	0.00000 (< 0.001 ***)	200	⌵
2	radio	0.332	0.329	0.00000 (< 0.001 ***)	200	⌵
3	newspaper_bin	0.106	0.064	0.00982 (< 0.01 **)	200	⌵

Y: sales X: newspaper Subgroup: none Facets: none

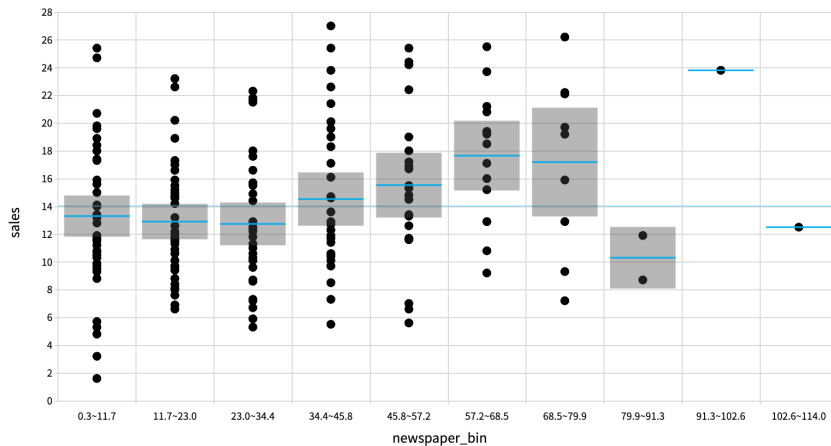
r: 0.23



Y: sales X: newspaper_bin

Subgroup: none Facets: none

⌵ ⌵ ⌵ ⌵ ⌵ ⌵





다정한 데이터 도구, HEARTCOUNT

- HEARTCOUNT(하트카운트)는 비전문가도 쉽게 엑셀 데이터셋을 업로드하여 시각화하고 분석할 수 있는 SaaS 솔루션입니다.
- Google 계정만 있다면, 홈페이지에서 바로 사용을 시작할 수 있어요!



다정한 데이터 도구, HEARTCOUNT

- 특장점으로는 '개별 레코드 수준의 시각화', '파생 변수 자동 생성', '패턴 자동 발견', '자연어 검색 + 설명' 등이 있습니다.

The screenshot displays the HEARTCOUNT interface with several key components:

- Left Panel (Filter & Analysis):** Shows a filter for '이억_bin' with 8,193 records. Below is a table of variable analysis results.
- Table of Variable Analysis:**

No.	변수명	R ²	Adjusted R ²
5	제품_분류 and 할인율_bin	0.074	0.072
6	제품대분류 and 할인율_bin	0.068	0.067
7	제품대분류	0.052	0.051
8	수량_bin	0.042	0.041
- Center Panel (Group Characteristics):** Displays characteristics for '그룹 A' and '그룹 B'.

변수명	비율
이억_percentile	82.69%
제품소분류	78.73%
제품대분류	48.62%
수량_percentile	40.74%
수량	31.64%
- Right Panel (Time Series Chart):** A stacked area chart showing '매출' (Sales) over time from 2014 to 2015. The chart includes a legend for '시간 간격 설명' (Time Interval Description) with options: 차분 (Difference), 년 (Year), 분기 (Quarter), 월 (Month), 일 (Day), 시 (Hour), and 없음 (None). The '주' (Week) option is selected.
- Far Right Panel (Settings):** A sidebar with settings for '레포트' (Report), '선택' (Select), 'Y축' (Y-axis), 'X축' (X-axis), '변수명' (Variable Name), and '배출방식' (Output Method).

학습하고 소통하는 공간, DATA HERO



- 데이터의 기초부터 실전까지, 전용 페이지에서 무료로 학습 가능
- 하트카운트팀은 물론 다양한 실무자들과의 소통 공간
- 다양한 집중 교육 캠프, 오프라인 밋업 등 이벤트

EDA(데이터 시각화) 강의

DATA HERO ORIGINAL CONTENTS

커뮤니티 소식

강의 VOD
EDA(데이터 시각화) 고급 통계 분석

실용 예제
EDA(데이터 시각화) 고급 통계 분석

블로그(아티클)
데이터나 멤버 오픈 스페이스

Upcoming Events

6월 웨비나 | 분석하기 좋은 데이터셋을 구성하는 요소를 6월 30일 (금) 오후 3:00 - 3:30 사전 등록 하기

Data Literacy

© Literacy, Numeracy, Data Literacy: 데이터 리터러시 에 대해 이해하기

데이터 분석 준비

EDA 101 (1): 분석하기 좋은 데이터셋, 변수 유형별 시각화 방법

데이터의 분포

EDA 101 (2): 데이터의 모양 묘사하기 (히스토그램, boxplot, percentile)

시각화 기초 문법

EDA 101 (3): 평균의 함정, 시각화 기본 문법, 상관계수 분석

DATAHERO # 3. 질문-답변 (데이터 히어로 커뮤니티 관련 궁금한 점은 이 곳에 남겨주세요. 778)

오전 4:52

오늘 웨비나에서 연사님이 DA가 일부 DE 업무까지 하는 경우??? 특정한 직무이름을 말씀 주셨었는데... 정확하게 떠오르지 않아서 질문드립니다!

오전 11:35

안녕하세요. 밋업 참여해서 강의 잘 듣고 있습니다. 혹시 물어 사칭이 있었나요? 저는 관심자과 물어 지체가 생경해서 이해도가 떨어지는 것 같아서요! 맥락상 이해는 하고 있지만 부족한 것 같이 느껴요. 있다면 공유 부탁드립니다.

#3. 질문-답변에 메시지 보내기



실습 시간

- EDA 도구: <https://www.heartcount.io/login>
- 데이터셋: [구글 "샘플 데이터셋" → 하트카운트 샘플 데이터셋 → 회귀분석 데이터셋](#)