국내 문학 신간 서적의 판매량 예측 모델 개발

- 선형 회귀분석 기법 활용 -

> **분석5기4조** <- data.frame(이름 = c("나성호", "고민정", "박대건", "최태웅", "유성용"),
>
> 호칭 = c("**K**evin", "**P**eter", "**D**avid", "**C**amel", "**H**arvey"))

목차

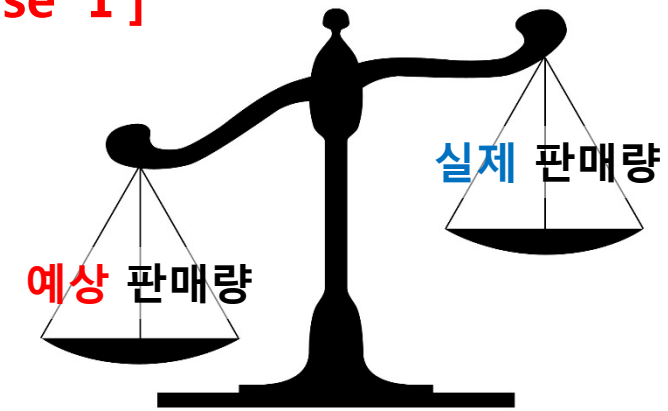
나는 궁금합니다!!

신간 서적이 나오면
이 책이 얼마나 팔릴지
미리 예측해야 하는데…
좋은 방법이 없을까요??

대형 온라인서점 Y사의
국내문학 상품기획자(MD)

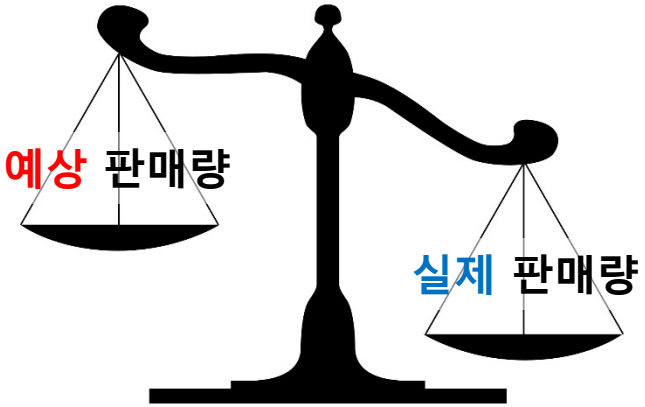
그래서 물었습니다. 왜 필요하신가요?

[Case Ⅰ]



- ✓ 판매 전, 출판사는 기분 좋고 MD의 어깨는 으쓱해지나,
- ✓ 판매 후, 월말 **악성재고**로 남아 **매몰비용**이 증가됨

[Case Ⅱ]



- ✓ 판매 전, 출판사는 마음 상하고 MD는 민망해질 뿐만 아니라,
- ✓ 판매 후, **영업기회 상실**에 따라 이 증가됨

MD의 요구사항

Y
MD

국내문학 신간 서적의 판매량을 가능한
정확하게 예측하고 싶습니다.

구체적인 판매량을 알고 싶으신가요?
아니면 **예상 등급** 정도면 될까요?

KPDCH

Y
MD

구체적인 수량으로 알려주시면 제가 예상
등급을 만들어 사용할 수 있을 것 같아요.

그렇다면, **회귀분석**으로 해야 되겠네요.
정확도는 대략 80% 정도면 될까요?

KPDCH

Y
MD

Data를 확보하라! (브레인스토밍)

텍스트 마이닝도 하죠.
네이버 책 사이트에는
대형 온라인 서점들의
베스트셀러 정보가
제공되고 있어요.

종속변수부터 확정하죠!
신간서적이 출시된 후,
"14일간의 판매량"으로
결정하면 될 것 같네요.

Y사의 거래 데이터로
저자와 출판사별로 과거
출판된 서적의 개수와
누적 총 판매량 등의
파생 변수를 만들어보죠.

우리가 분석할 서적들은
2013년 한 해 동안
새로 출간된 **국내문학**
서적이 좋겠습니다.

필요한 독립변수로는,
우선 **서점이 보유**하고
있는 데이터 중에서
저자와 출판사 정보를
들 수 있겠습니다.

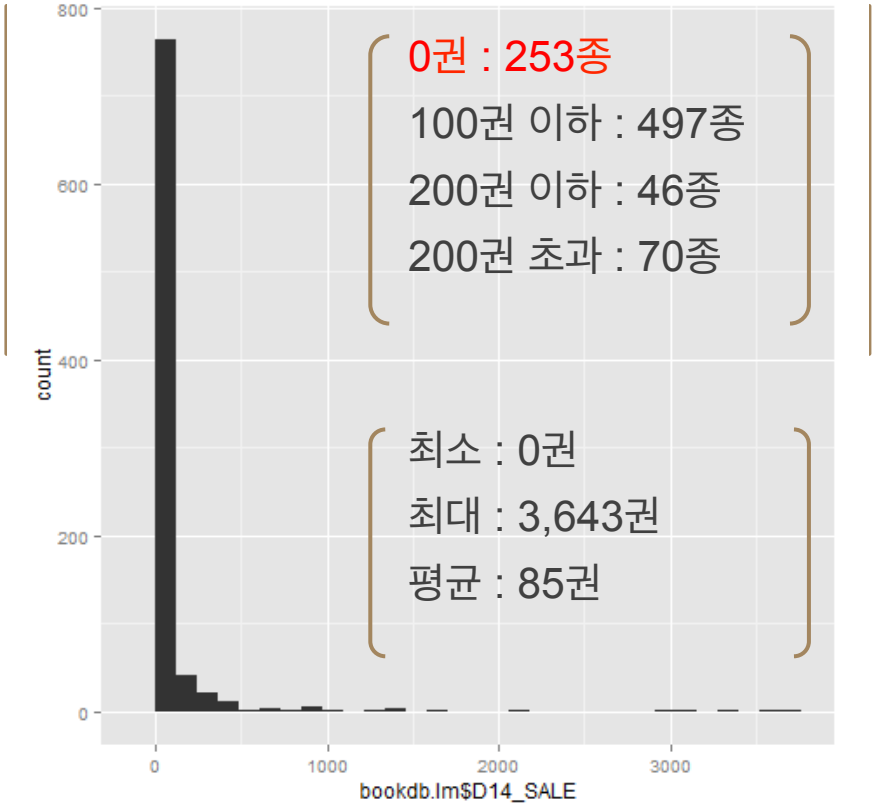
통계청 자료에 보면,
가구당 소득과 지출과
같은 통계량이 있는데,
가처분소득 정보를 새로
만들면 어떨까 싶네요~

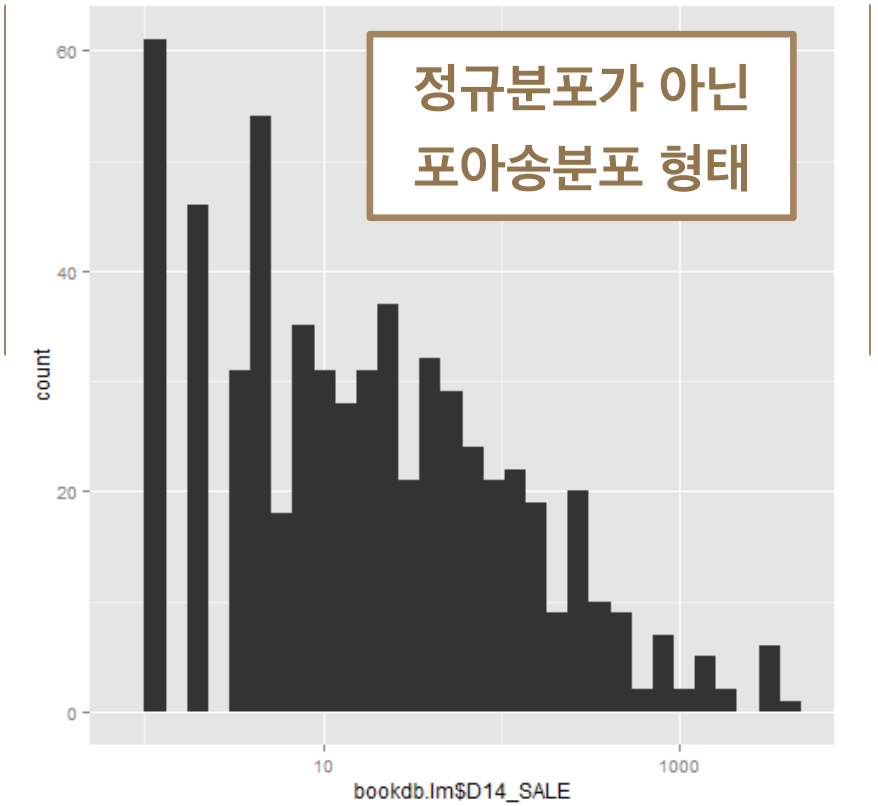
종속변수, 이렇게 생겼습니다.

종속변수의 정의

: **2013년 국내문학 신간 도서 866권**의
출간 이후 14일간 판매량 합계

종속변수에 log10을 씌워서 다시 그려 볼까요?





주요 독립변수의 생성

독립변수는 3가지 대분류로 나뉘며, 대분류별 주요 독립변수는 다음과 같습니다.

정형 데이터 (서점 DB)	비정형 데이터 (베스트셀러)	정형 데이터 (통계청)
▪ 저자 출판 종 수 ▪ **저자 총 판매량 (누적)** ▪ 저자 평균 판매량 ▪ 저자 최근 판매량 - 1M / 3M / 6M / 12M ▪ 출판사 출판 종 수 ▪ 출판사 총 판매량 ▪ 출판사 평균 판매량 ▪ 출판사 최근 판매량 - 1M / 3M / 6M / 12M ▪ 신간 서적 가격 (정가)	▪ **저자 베스트셀러 지수 (기간 중 BS 등록 횟수)** - 신간 서적의 출판일 전 1~4주간 등록 건 집계 - “네이버 책” 사이트에서 6대 온라인 서점 **데이터 크롤링**하여 확보 ▪ **장르별 판매 지수 개발** - 6대 서점의 장르별 BS 권 수 × ‘12년 M/S	▪ 가구당 소득 지수 - 분기별 가구당 소득/지출 - 분기별 가구당 서적 소비 ▪ 서적 판매량 - 월별 상품군별 판매액 - 월별 온라인몰 판매액 ▪ 월별 출판업 매출 전망 ▪ 월별 실업률

주요 독립변수의 생성 ① 정형 데이터 (서점 DB)

온라인 서점 Y사로부터 받은 원본 데이터로 몇 가지 파생변수를 만들었습니다.

ISBN(13)	도서명	출간일자	작가번호	출판사코드	정가	분야번호
9788994…	별을 사랑	20130615	1556	100634	6000	001006
9788997…	이어령의	20130506	1703	107513	45000	001006
9788954…	그 여자의	20130114	51032	6	11000	001006
9788932…	견딜 수	20130115	51036	79	8000	001006
:	:	:	:	:	:	:

- ✓ 전체 출판 종수 : 20만 여권
- ✓ '13년 일 판매량 : 100만 여건
- ✓ 작가 수 : 7만 여명
- ✓ 출판사 수 : 2만 여개

ISBN(13)	날짜	판매량
9788994…	20130615	1
9788994…	20130616	3
9788994…	20130617	2
9788994…	20130618	2
:	:	:

작가번호	출판 종수	누적판매량
1556	2	4301
1703	45	20054
51032	13	2435
51036	7	521
:	:	:

출판사코드	출판 종수	누적판매량
100634	302	10296
107513	1745	242154
6	25	7568
79	61	14325
:	:	:

주요 독립변수의 생성 ② 비정형 데이터 (베스트셀러)

네이버 책 사이트에서 6대 온라인 서점의 베스트셀러 정보를 크롤링하였습니다.

- ✓ 온라인 서점명
- ✓ 날짜 정보(년/월/주차)
- ✓ 베스트셀러 순위 (1~100위)
- ✓ 도서명
- ✓ 저자명
- ✓ 출판사명
- ✓ ISBN(13)
- ✓ 장르명

독립변수는 어떻게 선택하면 좋을까요?

준비된 **독립변수는 총 42개** 입니다.
보통 회귀분석 모형을 만들 때, 가용한 독립 변수들을 모두 포함하여 **stepwise 방식**으로 돌리면 변수를 자동으로 선택해주던데요.

먼저 종속변수와 독립변수간 상관관계를 확인해보세요.
유의확률 0.05 이하인 변수를 선택하시는 게 좋습니다.

정형 데이터 (서점 DB)

독립변수	상관계수	P-value
저자 출판종수	0.165	$9.5e^{-07}$
저자 총판매량	0.433	2.
저자 평균판매량	0.408	2.
출판사 출판종수	0.080	0.
출판사 총판매량	0.148	1.
출판사 평균판매량	0.263	3.
가격(정가)	-0.021	0.
최근 12M 판매량	0.280	$2.2e^{-16}$
:	:	:

비정형 데이터 (베스트셀러)

독립변수	상관계수	P-value
판매지수(경영)	-0.027	0.436
출판사지수(4w)	0.240	$7.6e^{-13}$
:	:	:

비정형 데이터 (통계청)

독립변수	상관계수	P-value
가구 총소득(Q)	-0.037	0.270
소비(Q)	-0.034	0.452
비(Q)	-0.012	0.725
리지수(M)	-0.014	0.680
상지수(M)	0.004	0.898
판매액(M)	-0.012	0.732
판매액(M)	-0.032	0.342
실업률(M)	-0.015	0.664
:	:	:

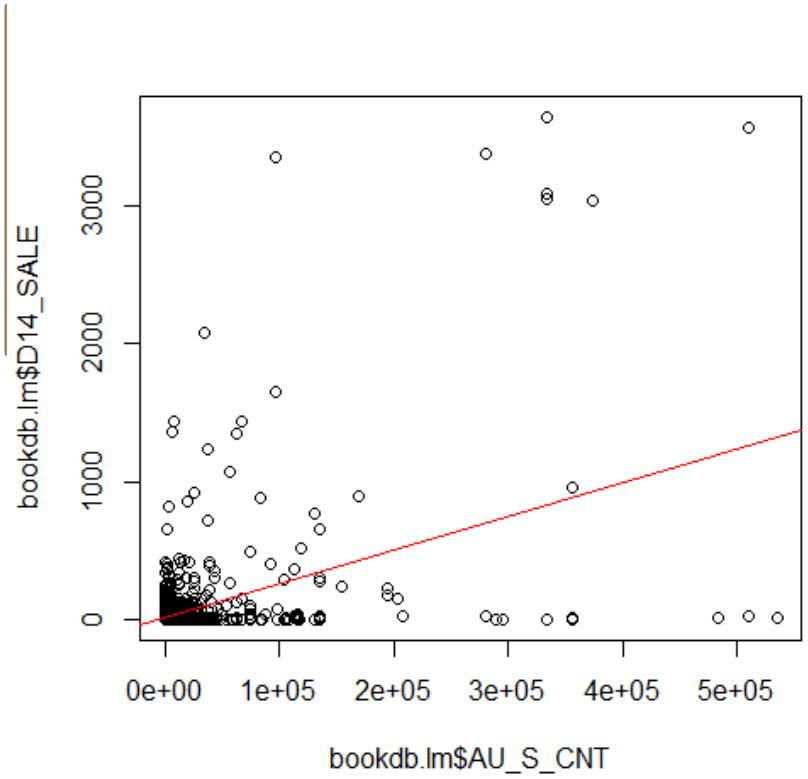
사용 가능 독립변수의 개수

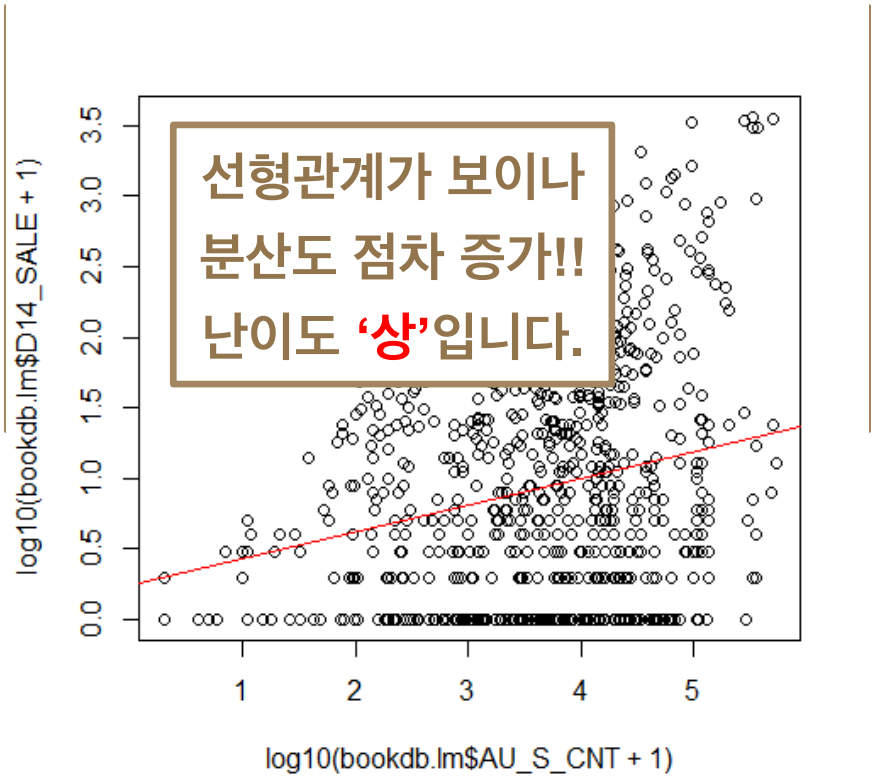
42개 → 24개

선형관계를 확인해봅시다!

정형 데이터 중 종속변수와 상관계수가 큰 **"저자 총 판매량"** 데이터로 선형관계를 확인해 보았습니다.
날씬한 직선일수록 강한 "선형관계"입니다.

종속변수와 독립변수 앞에 log10을 씌우고 다시 해보죠.
종속변수 값에 '0'이 있으니 '+1'을 하는 방법이 있습니다.





무한 삽질의 시작, 회귀분석을 하다 #1

먼저, 모델링과 검증을 위해 전체 데이터를 **Training Data(80%)**와 **Test Data(20%)**로 나눈 후, 24개 독립변수를 가지고 **stepwise** 방식의 다중회귀분석을 돌려보았습니다.

독립변수가 2개 이상인 다중 회귀분석인 경우 모델 적합도 확인은 Adjusted R^2로 해야죠. 이 모델은 0.7769이네요.

```
# linear model
> reg <- D14_SALE ~ .
> lm.tr <- step(lm(reg, data=bookdb.tr),
  direction="backward"))
> summary(lm.tr)
```

```
Residual standard error: 169.6 on 676 degrees of freedom
Multiple R-squared:  0.7808, Adjusted R-squared:  0.7769
F-statistic: 200.6 on 12 and 676 DF,  p-value: < 2.2e-16
```

Console D:/bigdata/project/data/

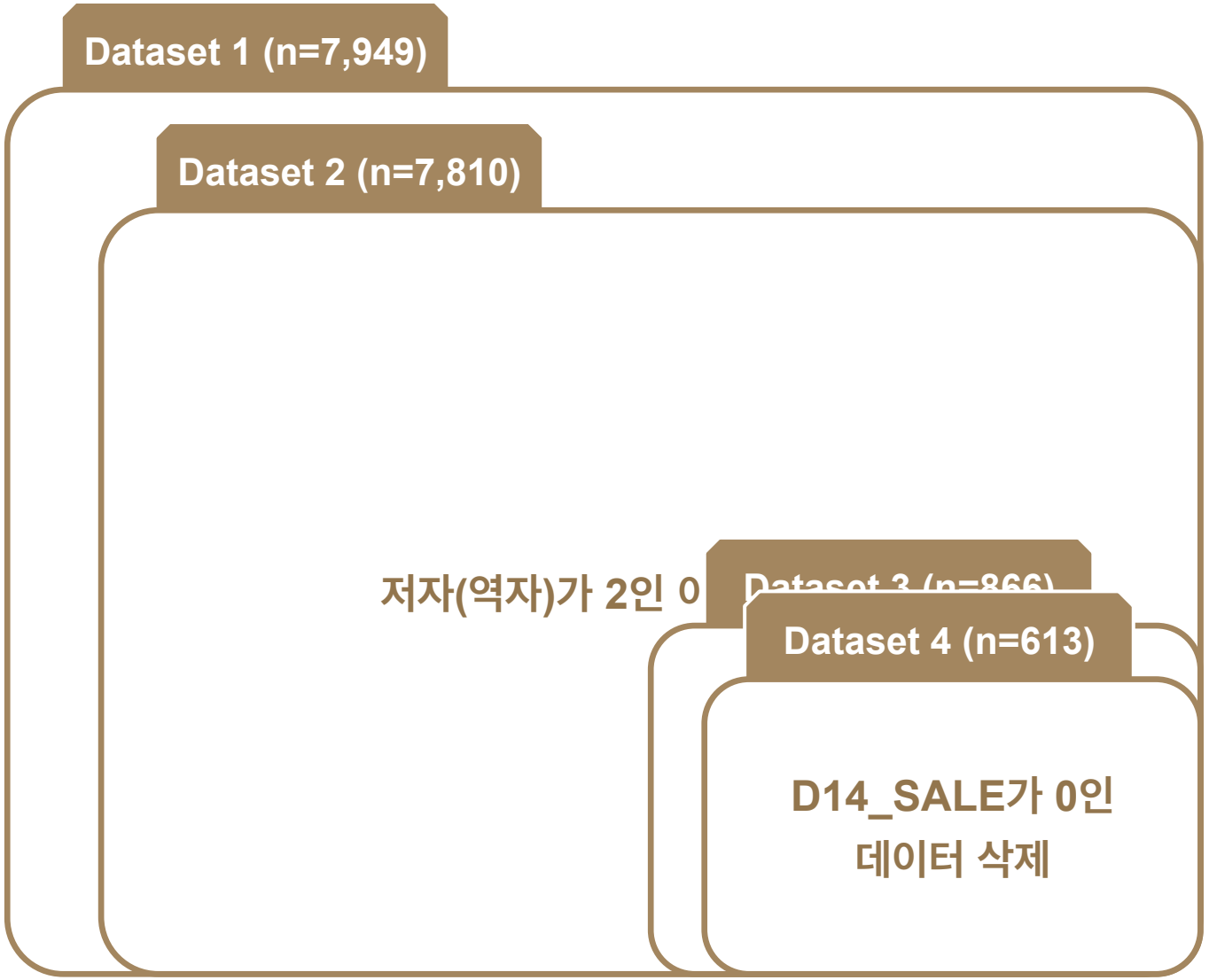
```
Call:
lm(formula = D14_SALE ~ AU_W_CNT + AU_S_CNT + PU_S_CNT + AU_BS_W1 +
    AU_BS_W2 + AU_BS_W3 + AU_BS_W4 + PU_BS_W4 + AU_S_12M + AU_S_03M +
    PU_S_12M + PU_S_03M, data = bookdb.lm.tr)

Residuals:
     Min       1Q   Median       3Q      Max
-1138.39   -23.76   -12.17    14.05  1434.67

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.386e+01  8.887e+00   2.685 0.007434 **
AU_W_CNT    -3.037e-01  1.183e-01  -2.567 0.010466 *
AU_S_CNT     2.180e-03  2.411e-04   9.041  < 2e-16 ***
PU_S_CNT    -4.634e-05  3.234e-05  -1.433 0.152407
AU_BS_W1    -6.517e+01  8.742e+00  -7.455 2.77e-13 ***
AU_BS_W2     1.289e+02  1.605e+01   8.028 4.40e-15 ***
                                      42  < 2e-16 ***
                                      39 0.000886 ***
                                      43 0.065787 .
                                      38 0.025573 *
                                      82 6.37e-09 ***
                                      58 0.031263 *
                                      76 0.076195 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 169.6 on 676 degrees of freedom
Multiple R-squared:  0.7808, Adjusted R-squared:  0.7769
F-statistic: 200.6 on 12 and 676 DF,  p-value: < 2.2e-16
```

[참고] 회귀방정식 탐색 과정 (Data Selection)

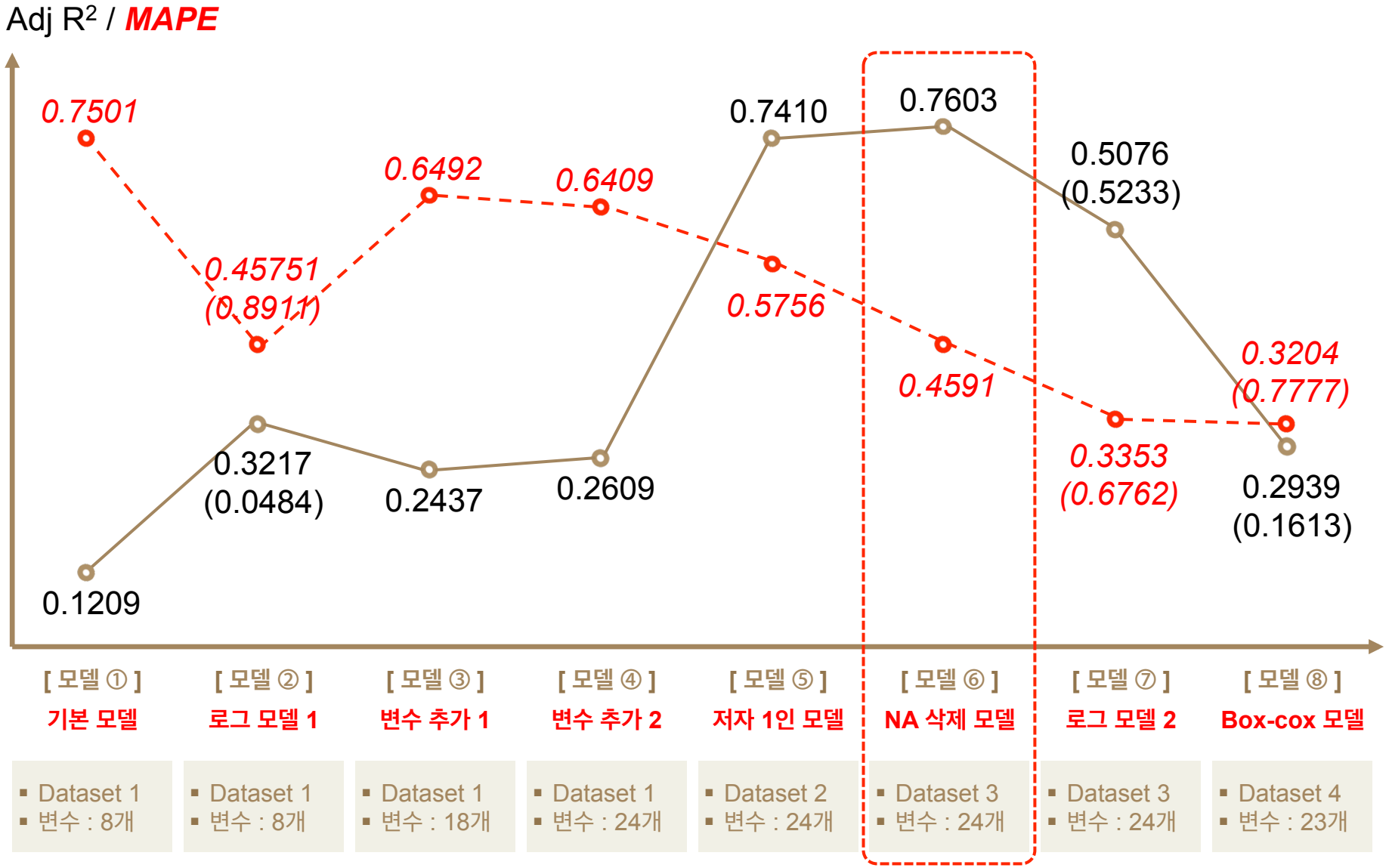


- **Variables group 1**
  : Y사 정형 데이터, 저자 및 출판사의 누적 출판 종 수, 누적/평균 판매량 등 8종

- **Variables group 2**
  : 저자 및 출판사 베스트셀러 지수 10종 추가

- **Variables group 3**
  : Y사 정형 데이터, 저자 및 출판사의 최근 1/3/6/12M 판매량 8종 추가

[참고] 회귀방정식 탐색 과정 (Model Selection)



무한 삽질의 시작, 회귀분석을 하다 #2

이번엔 종속변수의 분포에 맞는 일반화선형모델(GLM)을 검토해 볼까요? 먼저 Y값의 평균과 분산을 확인해 보시죠.

```
# mean & variance of y
> mean(bookdb.tr$D14_SALE)
[1] 84.77021

> var(bookdb.tr$D14_SALE)
[1] 116558.7
```

종속변수의 평균과 분산을 구해보니, 평균에 비해 분산이 매우 크네요. 이것으로 무엇을 알 수 있나요???

```
# GLM, poisson regression
> glm.pos <- glm(reg, data=bookdb.tr,
  family=poisson(link=log)))
> summary(glm.pos)

# GLM, Negative Binomial regression
> glm.nbr <- glm.nb(reg, data=
  bookdb.tr, method="glm.fit", link=log))
> summary(glm.nbr)
```

GLM에는 다양한 링크함수가 존재하는데요. 평균과 분산이 같을 땐 **poisson**을, 평균보다 분산이 매우 클 때는 **Negative Binomial**을 적용합니다.

무한 삽질의 시작, 회귀분석을 하다 #3

일반화선형모델(GLM) 2가지 모델을 만들어 summary를 살펴보니 Adjusted R^2를 확인할 수 없네요.

모델 적합도는 무엇으로 비교할 수 있나요?

보통 선형 모델간 적합도를 평가를 하기 위한 방법으로 MAPE(평균절대백분율오차)를 사용합니다.

 # Mean Absolute Percentage Error

실제값 대비 실제값과 예측값 편차 비율의 평균. 0에 가까울수록 모델 적합도 좋음

 # 모형 평가 지표

Max Error, 오분류율, Min/Max Error, sMAPE, MAE, RSS 등이 있음

선형모델간 적합도 비교 (w/ MAPE)

선형모델	MAPE
1. linear model	**0.4517**
2. GLM Poisson	0.4635
3. GLM Negative Binomial	0.7605

회귀분석에는 수많은 형태가 있지만,

어떤 모델이든 **쉽게 이해할 수 있고,**

적용이 편리한 모델을 사용하는 것이 정답!!

무한 삽질의 시작, 회귀분석을 하다 #4

앞서 몇 개 독립변수들의 회귀계수가 유의수준 0.05를 초과하는 경우도 있던데, 이들 변수들은 제거해야 하지 않을까요?
또 다중공선성 문제는 어떻게 해결하나요?

다중공선성이 높은 독립변수들을 먼저 제거하는 것도 좋은 방법입니다. 이 경우 VIF(분산팽창지수) 10 이상인 변수를 제외합니다.

```
# multicollinearity
> install.packages("alr3")
> library(alr3)
> vif(lm.tr)
```

```
# remove : PU_S_03M, ...
> lm.tr <- step(lm(reg2, data=bookdb.tr),
  direction="backward"))
> vif(lm.tr)
```

```
Console D:/bigdata/project/data/
> vif(lm.tr)
 AU_W_CNT  AU_S_CNT  PU_S_CNT  AU_BS_W1  AU_BS_W2  AU_BS_W3
 1.699444  4.708715 19.646536  3.847420  9.232303  7.155868
 AU_BS_W4  PU_BS_W4  AU_S_12M  AU_S_03M  PU_S_12M  PU_S_03M
 4.752288  2.898789  6.661455  4.822426 53.726169 25.177937
```

```
Console D:/bigdata/project/data/
> vif(lm.tr)
AU_W_CNT AU_S_CNT AU_BS_W1 AU_BS_W2 AU_BS_W3 AU_BS_W4
1.687116 4.612110 3.845024 9.208954 7.146605 4.700649
PU_BS_W4 AU_S_12M AU_S_03M
1.123384 5.959123 4.479529
```

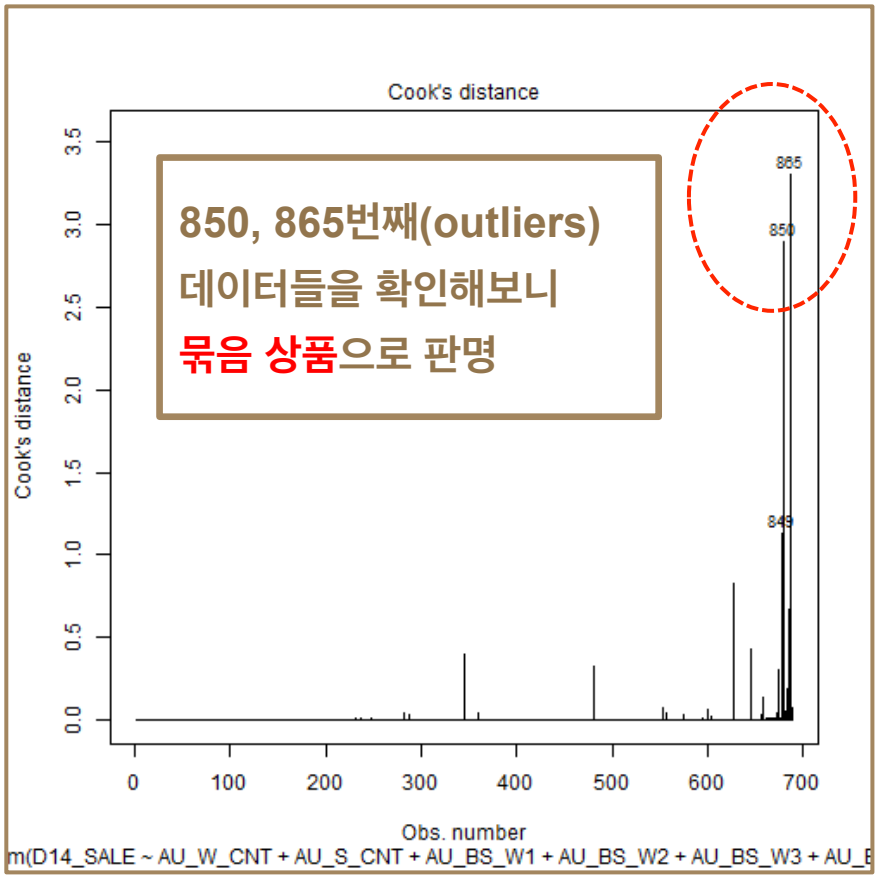
VIF가 10 이상인 독립변수를 제거할 때는
종속변수와의 상관계수가 가장 낮은 것부터
차례대로 삭제합니다.

새로운 회귀식의 모델 적합도

- Adj R^2 : 0.7769 → **0.7763** (-0.0006)
- MAPE : 0.4517 → **0.4479** (-0.0038)

무한 삽질의 시작, 회귀분석을 하다 #5

이제 다중공선성 문제도 해결했고,
회귀식과 계수들의 p-value들도 통계적으로
유의한 수준을 만족하고 있는 듯 합니다.



마지막으로 outliers가 있는지 확인해볼까요?
보통 Cook's Distance로 확인할 수 있습니다.

이상치를 제거한 회귀식의 모델 적합도

- Adj R^2 : 0.7763 → **0.8040** (+0.0277)
- MAPE : 0.4479 → **0.4476** (-0.0003)

테스트 데이터로 회귀모형을 검증한 결과

- Adj R^2 : **0.7624**
- MAPE : **0.4645**

회귀방정식의 완성

최종 회귀방정식을 정리하면 아래와 같습니다.

국내 문학 신간 서적의 출판 후 14일간 판매량

= Y절편 14.7

+ (0.0004 × 저자 총 판매량)

+ (-0.02 × 저자 최근 3M 판매량)

+ (-47.2 × 저자 베스트셀러 지수_w1)

+ (90.2 × 저자 베스트셀러 지수_w2)

+ (215.7 × 저자 베스트셀러 지수_w3)

+ (-30.3 × 저자 베스트셀러 지수_w4)

+ (7.84 × 출판사 베스트셀러 지수_w4)

회귀방정식의 해석

각 독립변수마다 측정 단위가 다르므로 **표준화 회귀계수를 통해 종속변수에 대한 설명력 비교**가 가능합니다!

```
# Standardized Beta Coefficient
> library(QuantPsyc)
> lm.beta(lm.tr)
```

AU_S_CNT	0.06335878
AU_BS_W1	-0.19753887
AU_BS_W2	0.32146338
AU_BS_W3	0.80222298
AU_BS_W4	-0.09416640
PU_BS_W4	0.07138233
AU_S_03M	-0.08274352

우리는 이 회귀방정식으로부터,

- ✓ **저자의 기존 서적들이 베스트셀러에 등록 될수록 신간 서적이 더 많이 팔리며,**
- ✓ **출판사보다 저자의 영향력이 더 크며,**
- ✓ **특히, 신간 서적의 출판일 3주 전에 저자가 베스트셀러로 등록된 경우, 판매량에 가장 큰 영향을 미치게 됨**을 알 수 있다.

종속변수에 영향을 크게 주는 변수의 순서

① 저자 베스트셀러 지수_3w

② 저자 베스트셀러 지수_2w

③ 저자 베스트셀러 지수_1w

KPDCH의 솔루션, 이렇게 활용해보세요.

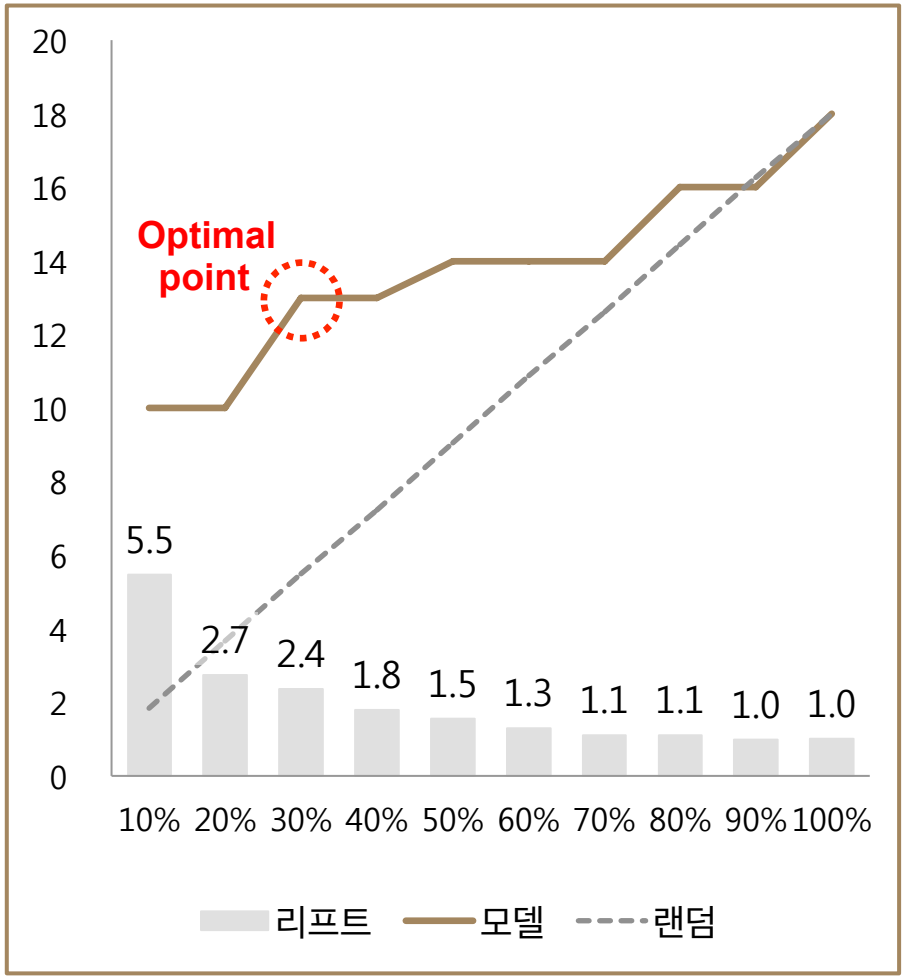
먼저 예측하고 싶은 신간 서적들을 대상으로
"출판 후 14일 판매량을 추정"해보세요.

[Test Data 회귀방정식 검증 결과]

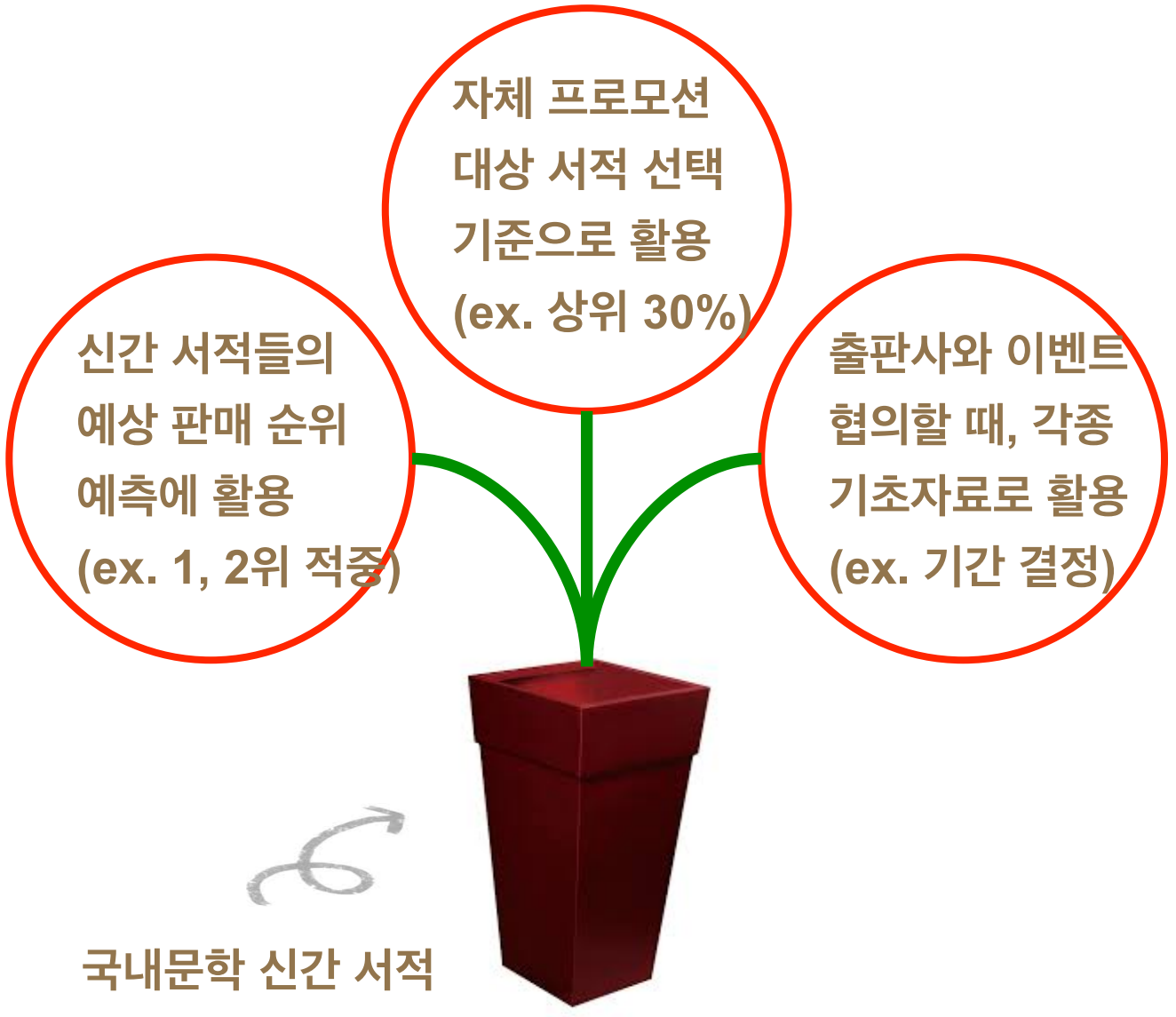
예측값 (누적)	1,000권 이상	100권 이상	10권 이상	1권 이상	1권 미만
종 수	6종	20종	162종	169종	177종
100권 점유비	66.7%	50.0%	9.9%	10.1%	10.2%

*** 1, 2위 서적의 순위를 정확하게 맞힘**

[ROC curve & Lift chart]



KPDCH의 솔루션, 이렇게 활용해보세요.



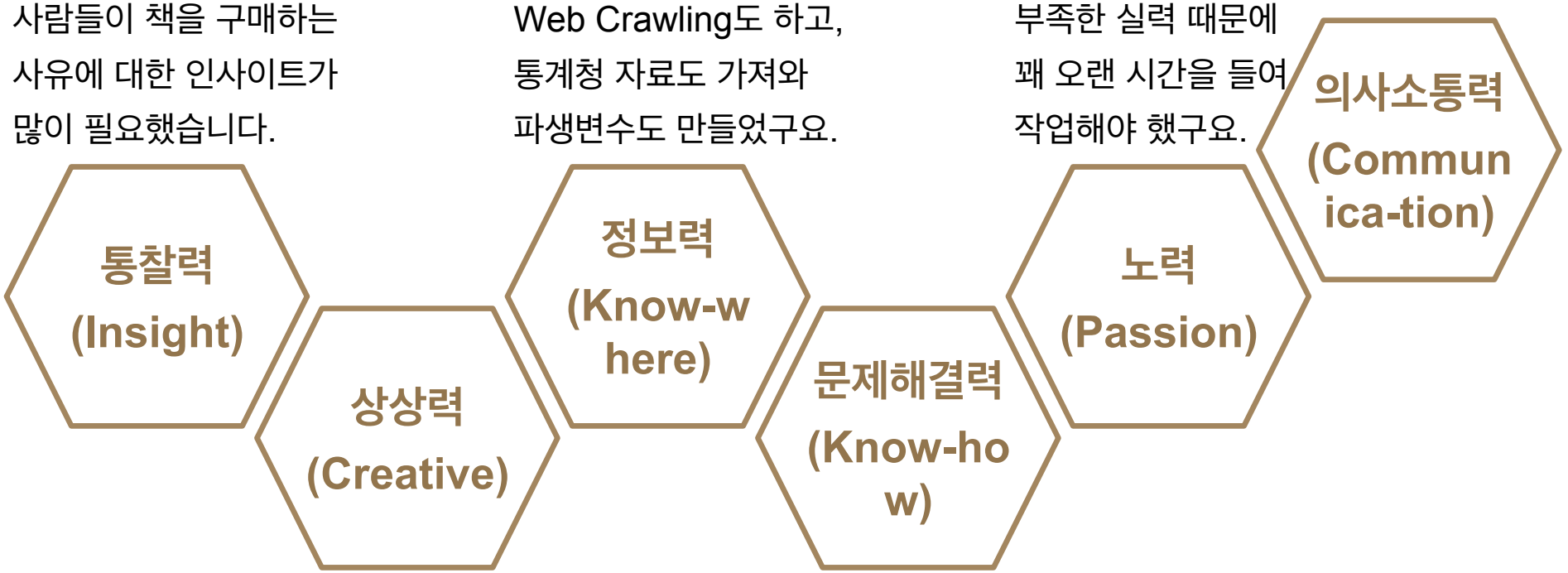
프로젝트를 마무리하며...

비록 결과는 미약하나,
그간의 과정을 이렇게
공유해 드립니다.

사람들이 책을 구매하는
사유에 대한 인사이트가
많이 필요했습니다.

Web Crawling도 하고,
통계청 자료도 가져와
파생변수도 만들었구요.

부족한 실력 때문에
꽤 오랜 시간을 들여
작업해야 했구요.



브레인스토밍을 통해
인사이트를 도출하고,
열심히 구글링도 했죠.

헌데, 회귀분석에 대해
제대로 알고 있는 게
그리 많지 않았습니다.

KPDCH

팀원 소개

유성용 주임	최태웅 연구원	박대건 대표	고민정 주임	나성호 수석
국민건강보험	㈜ 오픈메이트	케이에스비퓨처	WORDWORDS	하나금융 경영연구소
텍스트 마이닝	데이터 마트	데이터 마트	텍스트 마이닝	모델링

QnA

나성호
kevin.na74@gmail.com