가짜연구소

Causal-Lab

Part 2. 선형회귀

편향보정 - 유용한 선형회귀

2025.03.16 강지민

발표자 소개

강지민 (Jimin Kang)

- CJ ENM 커머스 부문 (2024.01~현재)
- 홈쇼핑, 모바일라이브, 이커머스 등 다양한 채널을 갖고 있는 커머스 플랫폼에서 데이터분석 담당

이번 파트에서 다루는 내용

Causal-Lab

- 편향 제거 기법으로써 선형회귀 분석의 필요성을 이해
- FWL 정리에 대한 개념 및 원리 이해

목차

선형회귀분석의 필요성

Causal-Lab

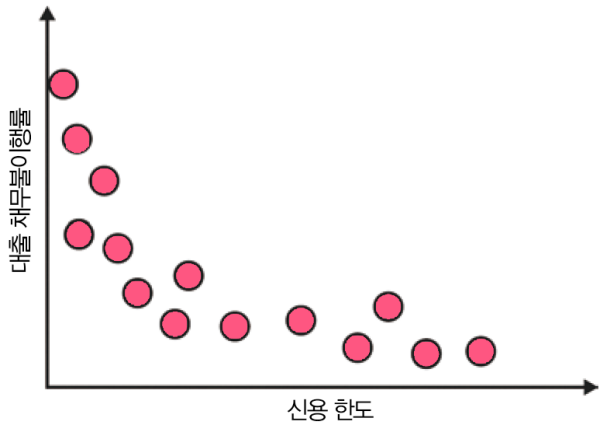
예시: 신용 카드 한도(T)가 채무불이행률(Y)에 영향을 미치는가?

- 일반적인 상식: 신용카드 한도 늘리면 채무불이행률이 높아질 것

선형회귀분석의 필요성

Causal-Lab

예시: 신용 카드 한도(T)가 채무불이행률(Y)에 영향을 미치는가?



- But, 실제 은행 데이터에서는 신용 한도(T)와 채무불이행률 (Y) 사이에 음의 상관관계가 나타남

  → 은행에서는 채무 불이행 가능성(Y)이 낮다고 판단되는 고객에게 더 높은 신용한도(T) 부여

  → 즉, 교란 편향의 영향 때문 (*교란 편향: T와 Y에 공통원인 때문에 발생하는 편향)

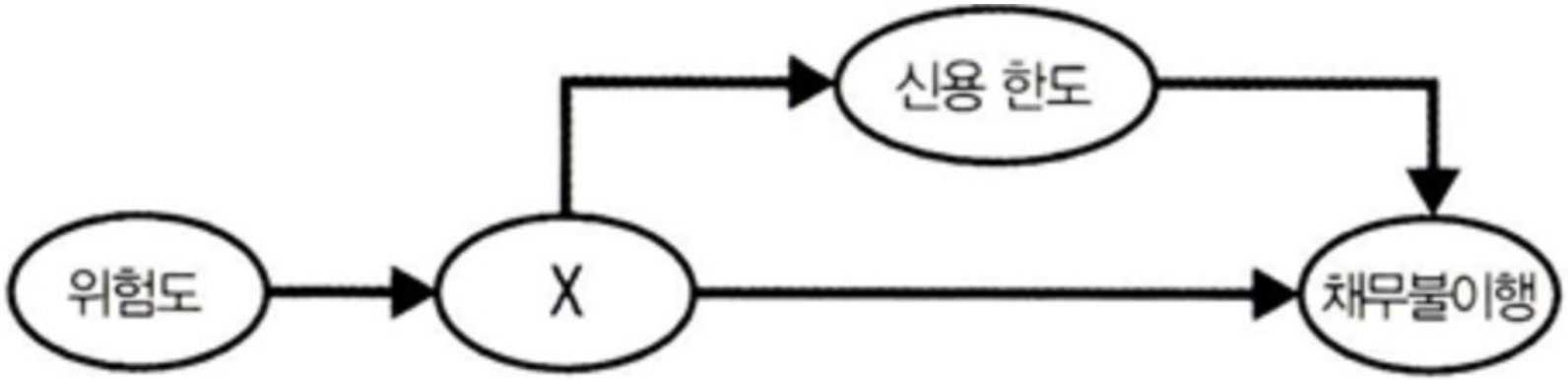
- 여기서 교란 편향은 T와 Y의 공통원인인 고객의 '위험도'

  → 교란을 줄 수 있는 변수를 잘 통제하는 것이 중요

선형회귀분석의 필요성

Causal-Lab

- T와 Y의 공통원인인 고객의 '위험도'를 알 수 없기 때문에, 대리변수 X로 이를 추정해보면 어떨까?
- 대리변수를 활용하여 T가 무작위 배정된 것처럼 보정할 수 있지 않을까?



선형회귀분석의 필요성

Causal-Lab

- 하지만 해당 방식은 그대로 적용하면, 오히려 상황이 나빠질 수 있다

- Why? 보정 공식을 활용해 변수를 보정하는 방법을 살펴보자

- $$ATE = E_x\{E[Y \mid T=1, X=x] - E[Y \mid T=0, X=x]\}$$

- 이 때, 특성 X가 많다면? 또 다른 문제 상황 발생 가능성, '차원의 저주'
  차원이 증가할수록 데이터가 고차원 공간에서 점점 멀리 퍼지게 되며,
  결국 모델이 학습할 패턴을 찾기 어려워짐→ 데이터 희소성(Data Sparsity) 증가

선형회귀분석의 필요성

Causal-Lab

- 이러한 차원의 저주 문제를 어떻게 해결할 수 있을까?

→ '선형회귀'

- X로 정의된 각각의 셀을 내삽(interpolate)하고 외삽(extrapolate)하는 것
- 결과 변수를 X변수로 투영한 후, 투영된 값을 바탕으로 실험군과 대조군 비교

선형회귀분석의 필요성

Causal-Lab

- 회귀추정식: $Default_i = \beta_0 + \beta_1 line_i + e_i$
- 추정된 β1^: 신용 한도가 1달러 증가할 때 채무불이행률이 얼마나 변할지에 대한 기댓값

- 은행은 위험이 적은 고객에게 더 높은 한도를 주는 경향이 있기 때문에, 교란 요인 보정 필요

- 회귀분석은 교란요인을 직접 보정하는 대신,
  OLS로 추정할 모델에 교란 요인 추가하여 그 영향을 분리하는 방식으로 해결

$$Default_i = \beta_0 + \beta_1 line_i + \theta_1 wage_i + \theta_2 creditScore1_i + \theta_3 creditScore2_i + e_i$$

교란 요인(임금, 신용점수 1, 신용점수 2) 추가

(* 장애모수 θ: 교란요인과 관련된 매개변수)

회귀분석 이론

Causal-Lab

- 선형회귀 분석의 목표: 평균제곱오차(MSE)를 최소화하는 최적의 회귀계수(β) 찾기

1) 단순선형회귀

$$\hat{\tau} = \frac{Cov(Y_i, T_i)}{Var(T_i)}$$

→ T와 Y의 공분산을 T의 분산으로 나눈 값

→ 의미: 변수 T가 1 증가할 때 결과 Y가 평균적으로 얼마나 변하는지

회귀분석 이론

Causal-Lab

- 선형회귀 분석의 목표: 평균제곱오차(MSE)를 최소화하는 최적의 회귀계수(β) 찾기

2) 다중선형회귀: 변수가 두 개 이상 일 때

$$y_i = \beta_0 + T_i\tau + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$$

$$\hat{\tau} = \frac{Cov\left(Y_i, \widetilde{T}_i\right)}{Var\left(\widetilde{T}_i\right)}$$

- 여기서 핵심은 T의 효과(τ)를 추정하는 것
- T가 Y에 미치는 순수한 영향을 분석하고 싶은데, 다른 변수들(X1, X2, ...)도 고려해야 함
- 다른 변수를 사용하여 T를 예측할 수 있다면, T는 무작위가 아니기 때문에

  → 다른 변수(모든 교란 요인) X를 통제하면, T를 무작위처럼 보이게 할 수 있음

회귀분석 이론

Causal-Lab

- 이때, 어떻게 교란 요인을 제거할까?

  (1) 선형회귀분석을 통해 교란 요인 X에서 T를 예측한다

  (2) T에서 해당 회귀에 대한 잔치인 T(~)를 빼준다

  → T를 예측하는데 이미 사용한 변수 X를 이용해서 T(~)를 예측할 수 없기 때문

- 즉, T(~)는 X의 다른 변수와 연관이 없는 버전의 또 다른 T가 되는 것

→ 이 부분이 "FWL(프리슈-워-로벨 정리)"

편향 제거 기법: FWL 스타일의 직교화

Causal-Lab

- 가장 먼저 사용할 수 있는 편향제거 기법으로 간단하면서도 강력함
- 비실험 데이터를 처치 T가 무작위 배정된 것처럼 보이게 함

- FWL 정리로부터 구한 결과와 다중회귀분석의 결과는 동일함

- FWL 정리는 3단계로 나누어 추정할 수 있고
  편향 제거와 잡음 제거한 데이터셋을 얻을 수 있음

편향 제거 기법: FWL 스타일의 직교화

Causal-Lab

Step 1. 편향 제거 단계 ⭐

- 처치 T를 교란 요인 X에 회귀하여 처치 잔차 T(~) = T-T^ 구하기

  → 이 단계는 처치가 다른 변수들에 의해 영향을 받는 정도(편향)를 제거하는 단계

Step 2. 잡음 제거 단계

- 결과 Y를 교란 요인 X에 대해 회귀하여 결과 잔차 Y(~) = Y-Y^ 구하기

  → 이 단계는 결과가 교란 요인에 의해 왜곡된 부분(잡음)을 제거

Step 3. 결과 모델 단계

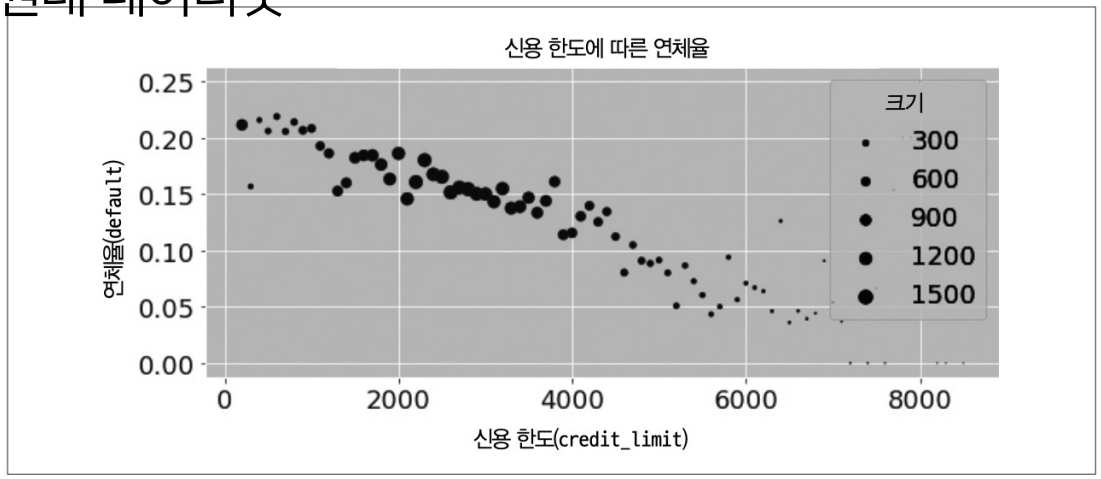
- 결과 잔차 Y(~)를 처치 잔차 T(~)에 회귀하여 T가 Y에 미치는 인과효과 추정값 구하기

  →이 단계에서는 인과효과를 정확히 추정하는 것이 목표

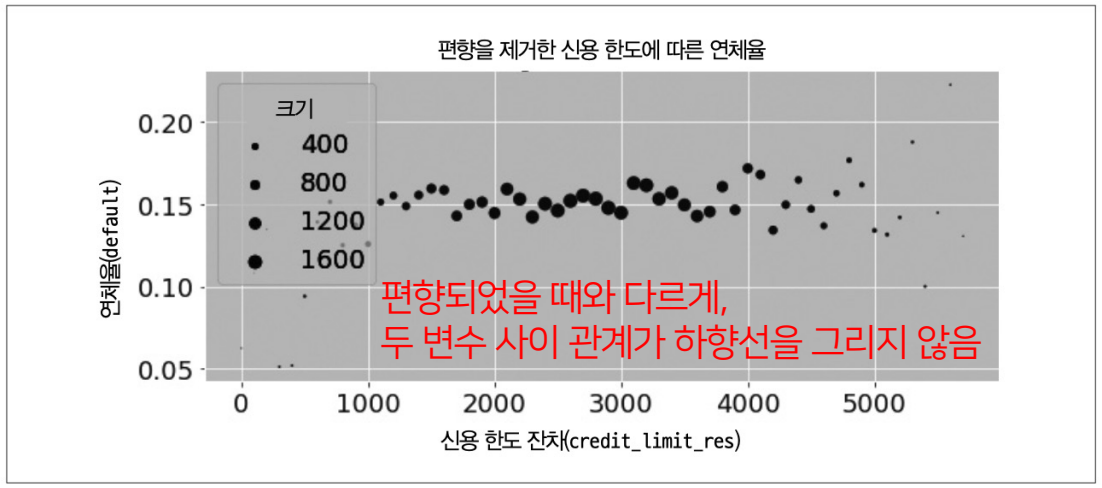
편향 제거 기법: FWL 스타일의 직교화

Causal-Lab

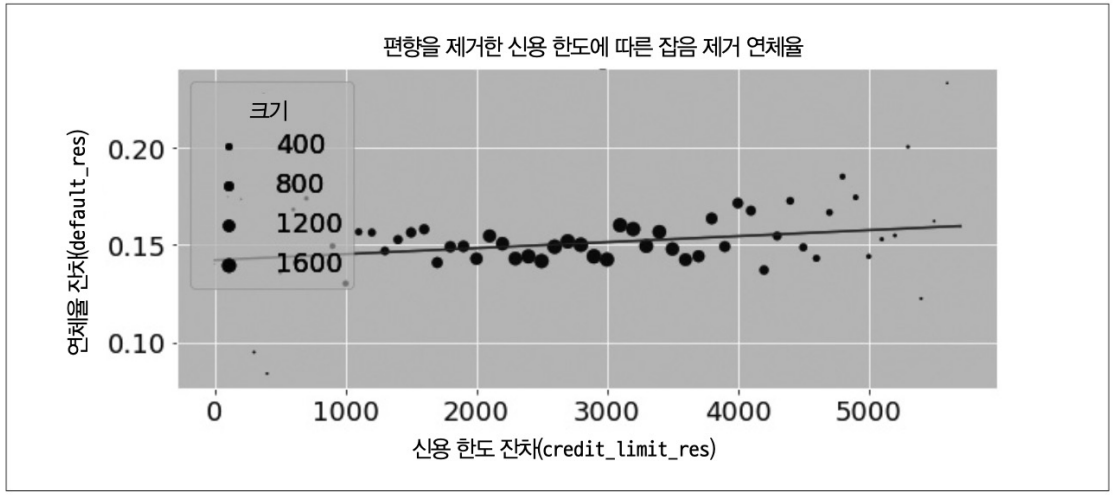
원래 데이터셋



1) 편향 제거 후 데이터셋



2) 잡음 제거 후 최종 결과 모델



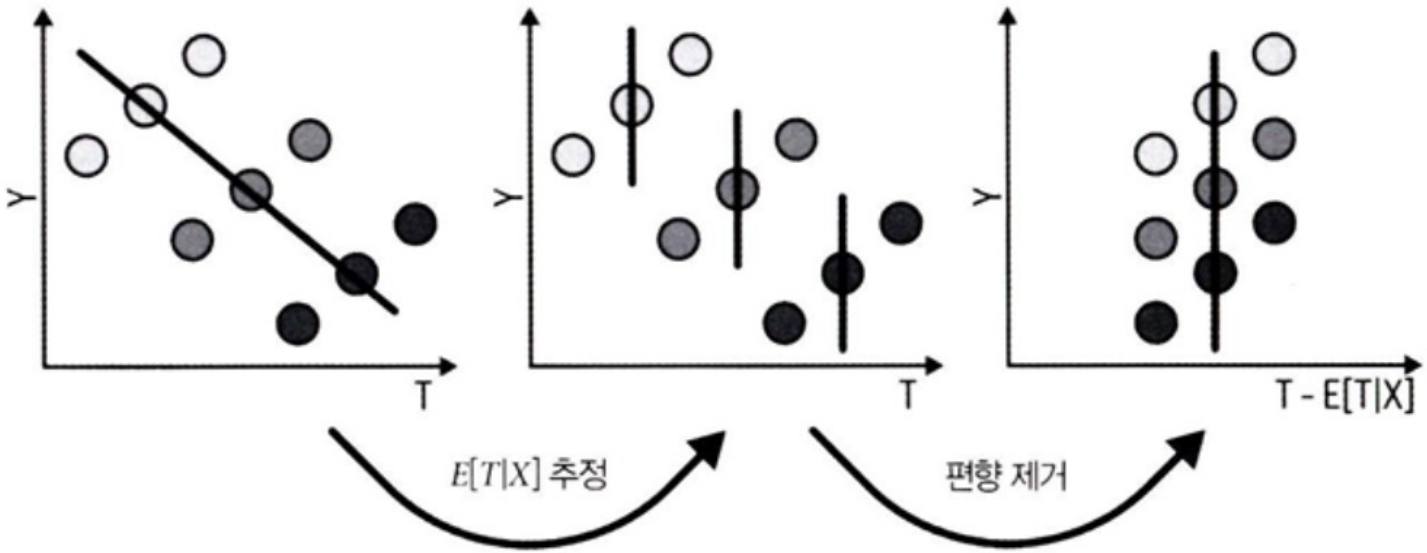
편향 제거 기법: FWL 스타일의 직교화

Causal-Lab

FWL 정리 시각화 요약

- T와 Y의 관계를 추정하고 싶지만 교란요인 X가 존재할 때

  (1) 편향 제거 단계에서 E[T|X]를 추정 (2) 편향이 제거된 T(~) = T – E[T|X] 구하기

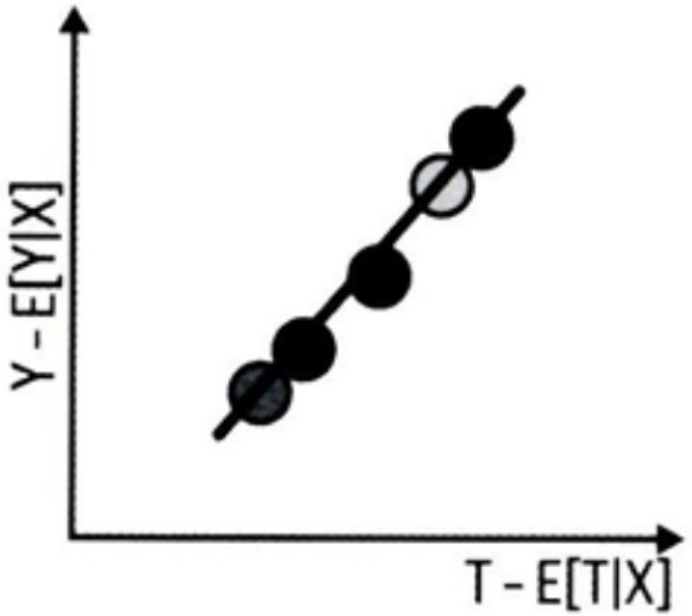


편향 제거 기법: FWL 스타일의 직교화

Causal-Lab

FWL 정리 시각화 요약

- 편향과 잡음을 제거한 후 T와 Y 사이의 양의 관계 확인 가능
- 이 회귀식은 Y를 T와 X에 동시에 회귀했을 때의 기울기와 정확히 같음



선형회귀에서의 비선형성

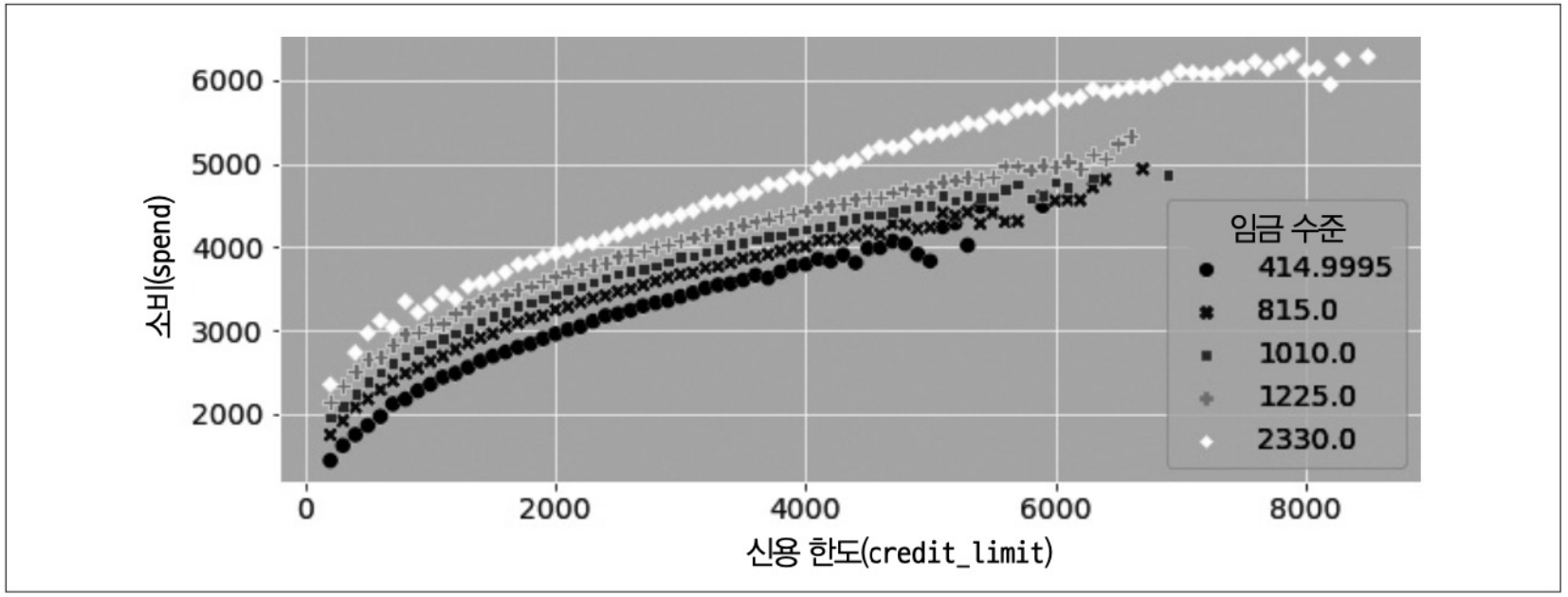
Causal-Lab

지금까지는 처치 반응 곡선이 선형적이었지만..

실제로는 선형적이지 않은 상황을 마주할 가능성이 더 높음!

예를 들어, 신용 한도를 2,000에서 3,000으로 늘렸을 때보다

1,000에서 2,000으로 늘렸을 때 소비가 더 많이 증가하는 경우 어떻게 해야할까?



선형회귀에서의 비선형성

Causal-Lab

비선형성을 해결하기 위해, 처치와 결과를 선형 관계로 변환 필요

- 고려할 수 있는 모델: 로그 함수, 제곱근 함수 등

함수를 선택할 때 고려해야 할 요소 '데이터 특성'

- 로그 변환 → 데이터가 매우 비대칭하게 분포하거나 값이 급격히 증가하는 경우
- 제곱근 변환 → 급격한 변화보다 점진적으로 증가하는 패턴을 보이는 데이터

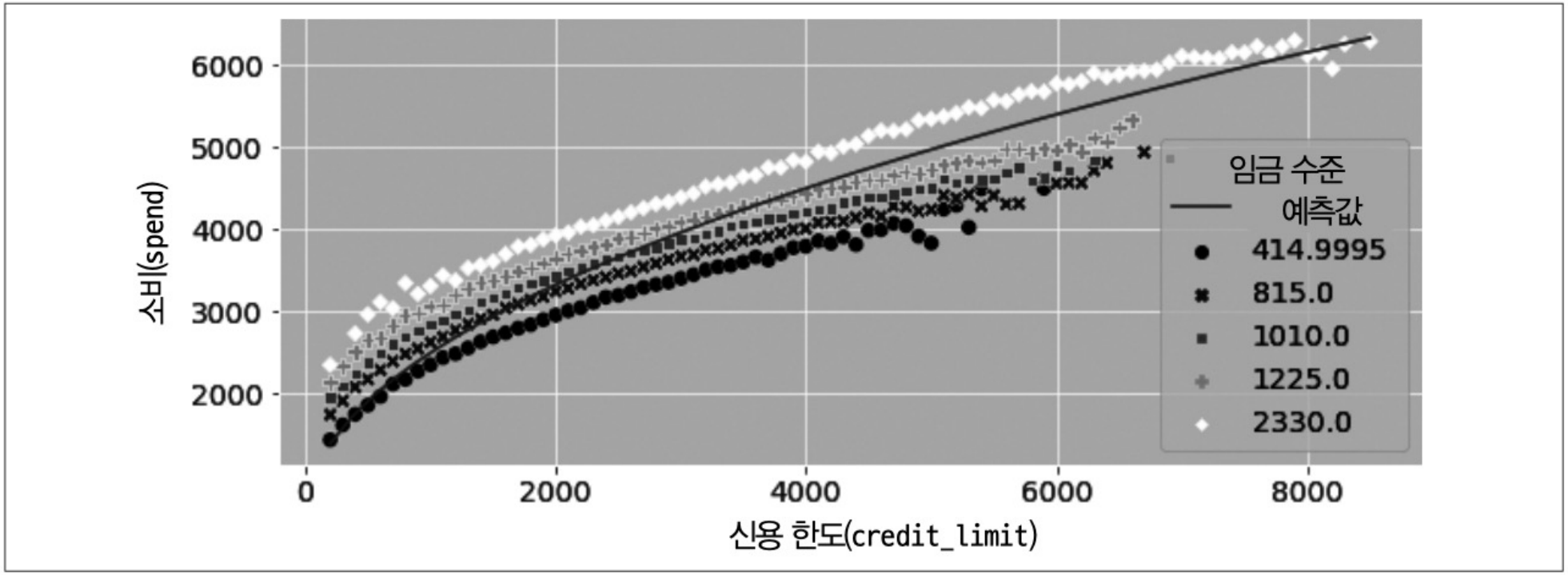
선형회귀에서의 비선형성

Causal-Lab

비선형성을 해결하기 위해, 처치와 결과를 선형 관계로 변환 필요

- 제곱근 함수를 통해 선형 관계로 변환

$$spend_i = \beta_0 + \beta_1 \sqrt{line_i} + e_i$$



비선형 FWL과 편향제거

Causal-Lab

Step 1. 처치 선형화 단계: T와 Y의 관계를 선형화하는 함수 F 찾기 (추가)

Step 2. 편향 제거 단계: F(T)를 교란 요인 X에 회귀하고 처치 잔차 F(T)~=F(T)−F(T)^구하기

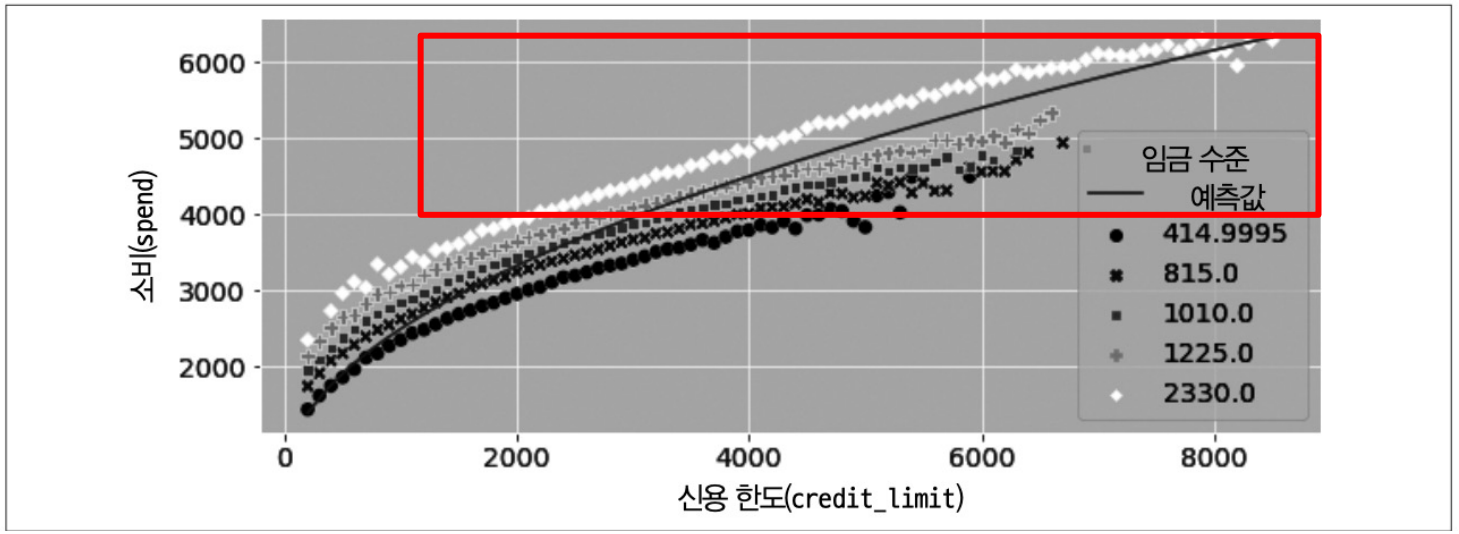
Step 3. 잡음 제거 단계: 결과 Y를 교란 요인 X에 대해 회귀하여 결과 잔차 Y~=Y−Y^ 구하기

Step 4. 결과 모델 단계: 이렇게 얻은 결과 모델은 결과 잔차 Y~를 처치 잔차 F(T)~에 대해 회귀하여 F(T)가 Y에 미치는 인과효과 추정값 구하기

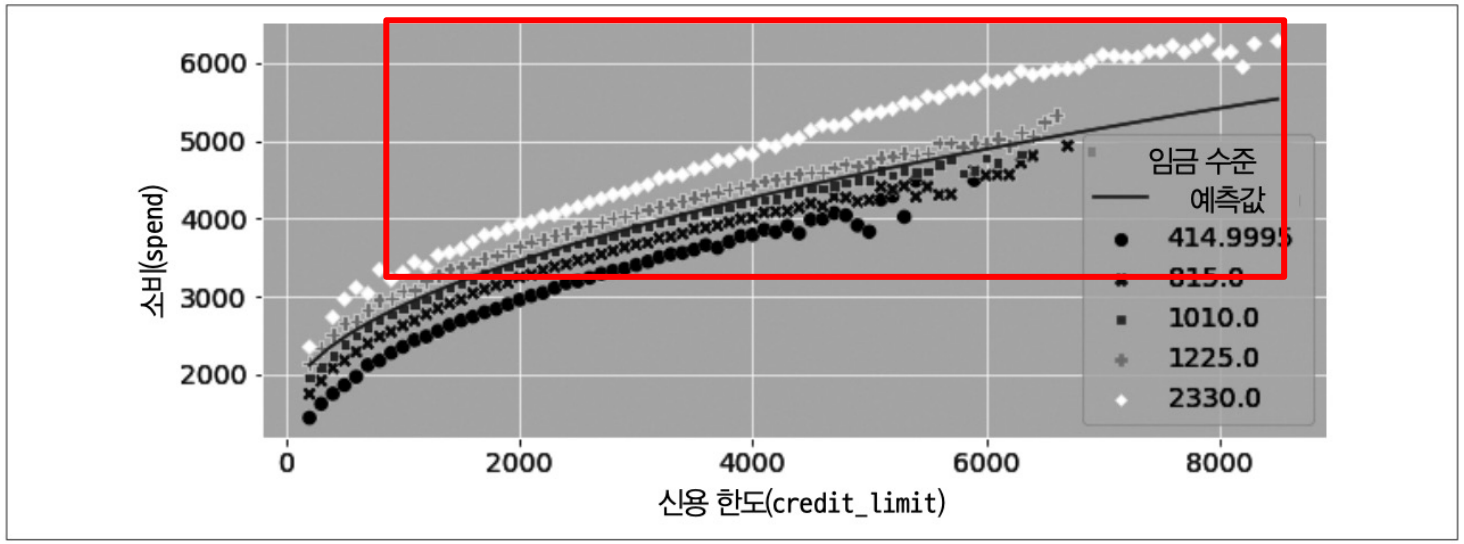
비선형 FWL과 편향제거

Causal-Lab

비선형 FWL 이전



비선형 FWL 한 결과



이전처럼 상향 편향되지 않고

임금 그룹의 중간을 정확히 통과

더미변수를 활용한 회귀분석

Causal-Lab

- 모델에 모든 교란 요인이 포함되는지 판단하기는 매우 어려움
- 그렇기 때문에, 가능하다면 무작위 실험을 하는 것이 좋으나 큰 비용이 발생함

→ 무작위 실험의 차선책으로 '조건부 무작위 실험'을 활용할 수 있음

조건부 무작위 실험

- 공변량 X에 따라 서로 다른 분포에서 표본을 뽑아 여러 국소 실험을 만드는 것
- 장점: 조건부 독립 가정에 더 설득력이 생김
- 단점: 실험군에 대한 결과만으로 단순 회귀분석을 하면 편향된 추정값을 얻게 됨

더미변수를 활용한 회귀분석

Causal-Lab

예시: 신용 카드 한도(T)가 채무불이행률(Y)에 영향을 미치는가?

- 교란 요인 포함하지 않고 모델 추정한 경우

$$default_i = \beta_0 + \beta_1 lines_i + e_i$$

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1369	0.009	15.081	0.000	0.119	0.155
credit_limit	-9.344e-06	1.85e-06	-5.048	0.000	-1.3e-05	-5.72e-06

- 더 높은 신용한도가 고객의 채무불이행 위험을 낮추지는 않으므로 추정값이 음수인 것은 말이 안 됨
- 실험 설계 방식으로 인해, credit_score1 가 낮은 고객이 평균적으로 더 높은 한도(T)를 받았기 때문에 음수값이 나온 것

더미변수를 활용한 회귀분석

Causal-Lab

이를 보정하기 위해 모델에 처치가 무작위로 배정된 그룹 정보 포함시켜야 함

→ 원-핫 인코딩으로 더미변수 만들기

- 더미변수는 그룹에 대한 이진값으로 구성 (그룹에 속하면 1, 아니면 0)

- 이번 예시에서는 대리변수 credit_score1_buckets를 통제
- 더미 변수 5개 생성하고 이를 포함시켜 B1을 재추정

	wage	educ	exper	married	...	sb_400	sb_600	sb_800	sb_1000
0	890.0	11	16	1	...	1	0	0	0
1	670.0	11	7	1	...	0	0	0	0
2	1220.0	14	9	1	...	1	0	0	0
3	1210.0	15	8	1	...	0	1	0	0
4	900.0	16	1	1	...	0	0	0	0

$$default_i = \beta_0 + \beta_1 lines_i + \theta \mathbf{G}_i + e_i$$

더미변수를 활용한 회귀분석

Causal-Lab

- 재추정 결과

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2253	0.056	4.000	0.000	0.115	0.336
credit_limit	4.652e-06	2.02e-06	2.305	0.021	6.97e-07	8.61e-06
sb_200	-0.0559	0.057	-0.981	0.327	-0.168	0.056
sb_400	-0.1442	0.057	-2.538	0.011	-0.256	-0.033
sb_600	-0.2148	0.057	-3.756	0.000	-0.327	-0.103
sb_800	-0.2489	0.060	-4.181	0.000	-0.366	-0.132
sb_1000	-0.2541	0.094	-2.715	0.007	-0.438	-0.071

- 더미변수 처치 후 credit_limit의 B1 추정량이 제대로 양수가 나온 모습
- 기울기 매개변수는 B1 하나로, 교란요인을 통제하려고 더미변수를 추가했을 때 절편은 그룹 당 하나씩 생기지만 모든 그룹에 동일한 기울기가 적용됨

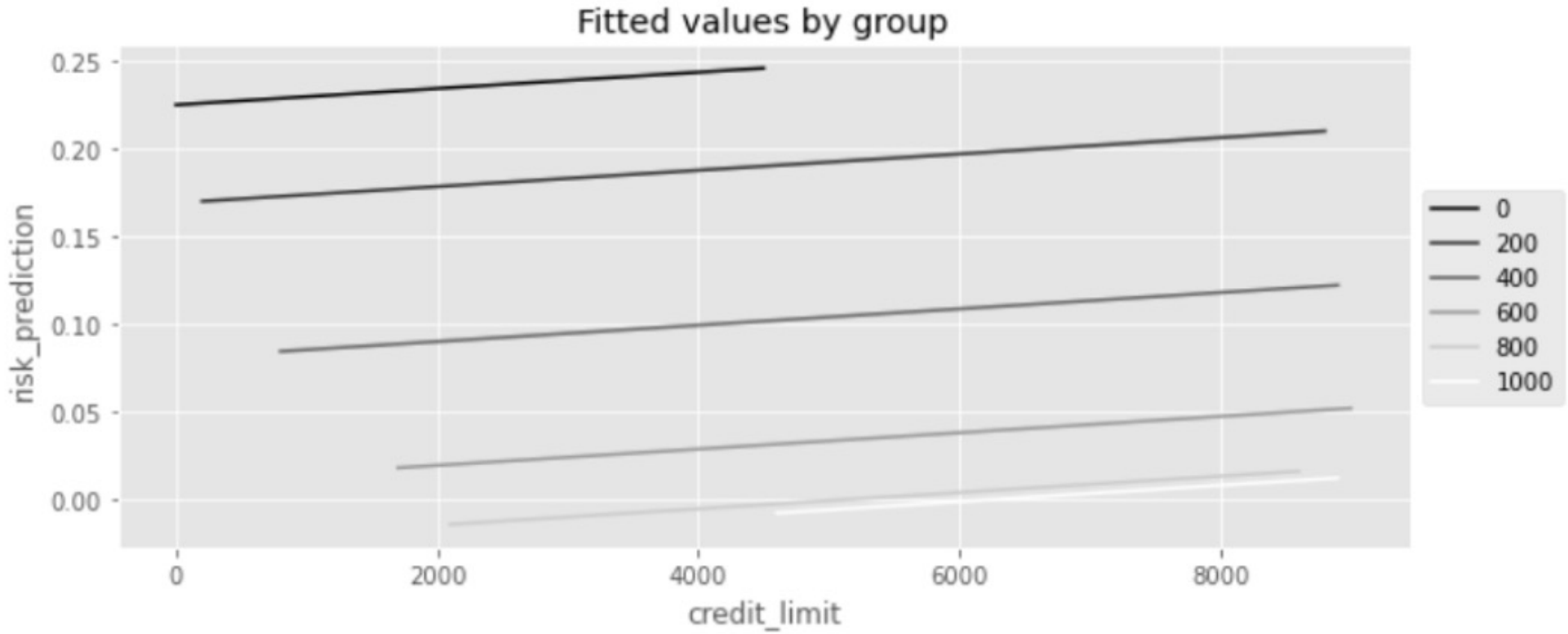
더미변수를 활용한 회귀분석

Causal-Lab

- 이 때 기울기 매개변수는 B1 하나로,

  교란요인을 통제하려고 더미변수를 추가했을 때

  절편은 그룹 당 하나씩 생기지만 모든 그룹에 동일한 기울기가 적용됨



포화 모델

Causal-Lab

- 포화 모델: 회귀분석에서 모든 가능한 설명변수와 그 상호작용을 포함하는 모델
- 주어진 데이터에 대해 모든 변수와 상호작용을 포함하므로 과적합 위험이 있음
- 그러나 각 더미변수에 해당하는 특정 그룹의 인과효과를 파악하는 데에 유용하게 쓰임

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3137	0.077	4.086	0.000	0.163	0.464
C(credit_score1_buckets)[T.200]	-0.1521	0.079	-1.926	0.054	-0.307	0.003
C(credit_score1_buckets)[T.400]	-0.2339	0.078	-3.005	0.003	-0.386	-0.081
C(credit_score1_buckets)[T.600]	-0.2957	0.080	-3.690	0.000	-0.453	-0.139
C(credit_score1_buckets)[T.800]	-0.3227	0.111	-2.919	0.004	-0.539	-0.106
C(credit_score1_buckets)[T.1000]	-0.3137	0.428	-0.733	0.464	-1.153	0.525
credit_limit	-7.072e-05	4.45e-05	-1.588	0.112	-0.000	1.66e-05
credit_limit:C(credit_score1_buckets)[T.200]	7.769e-05	4.48e-05	1.734	0.083	-1.01e-05	0.000
credit_limit:C(credit_score1_buckets)[T.400]	7.565e-05	4.46e-05	1.696	0.090	-1.18e-05	0.000
credit_limit:C(credit_score1_buckets)[T.600]	7.398e-05	4.47e-05	1.655	0.098	-1.37e-05	0.000
credit_limit:C(credit_score1_buckets)[T.800]	7.286e-05	4.65e-05	1.567	0.117	-1.83e-05	0.000
credit_limit:C(credit_score1_buckets)[T.1000]	7.072e-05	8.05e-05	0.878	0.380	-8.71e-05	0.000

중립 통제변수

Causal-Lab

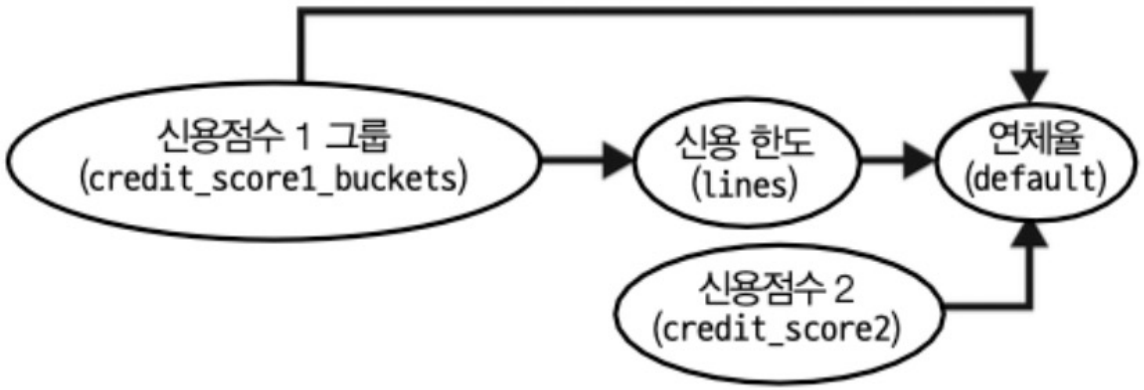
- 회귀분석에서 고려해야 할 다른 유형의 변수 '중립 통제변수'

  → 이러한 통제변수는 회귀분석 추정에서 편향에는 영향을 미치지 않음 (중립적)

  → 하지만, 분산에는 심각한 영향을 줄 수 있음

(회귀분석에서 특정변수를 포함할 때는 편향-분산 트레이드오프가 존재)

- 예시모델에 credit_score2를 포함하는 것이 좋을까?



중립 통제변수

Causal-Lab

- 모델에 credit_score2를 포함하는 것이 좋을까?

* credit_score2를 포함하지 않은 경우

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2253	0.056	4.000	0.000	0.115	0.336
credit_limit	4.652e-06	2.02e-06	2.305	0.021	6.97e-07	8.61e-06
sb_200	-0.0559	0.057	-0.981	0.327	-0.168	0.056
sb_400	-0.1442	0.057	-2.538	0.011	-0.256	-0.033
sb_600	-0.2148	0.057	-3.756	0.000	-0.327	-0.103
sb_800	-0.2489	0.060	-4.181	0.000	-0.366	-0.132
sb_1000	-0.2541	0.094	-2.715	0.007	-0.438	-0.071

* credit_score2를 포함한 경우

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5576	0.055	10.132	0.000	0.450	0.665
C(credit_score1_buckets)[T.200]	-0.0387	0.055	-0.710	0.478	-0.146	0.068
C(credit_score1_buckets)[T.400]	-0.1032	0.054	-1.898	0.058	-0.210	0.003
C(credit_score1_buckets)[T.600]	-0.1410	0.055	-2.574	0.010	-0.248	-0.034
C(credit_score1_buckets)[T.800]	-0.1161	0.057	-2.031	0.042	-0.228	-0.004
C(credit_score1_buckets)[T.1000]	-0.0430	0.090	-0.479	0.632	-0.219	0.133
credit_limit	4.928e-06	1.93e-06	2.551	0.011	1.14e-06	8.71e-06
credit_score2	-0.0007	2.34e-05	-30.225	0.000	-0.001	-0.001

- 표준오차가 감소 → 추가한 변수가 선형회귀의 잡음 제거 단계에 기여

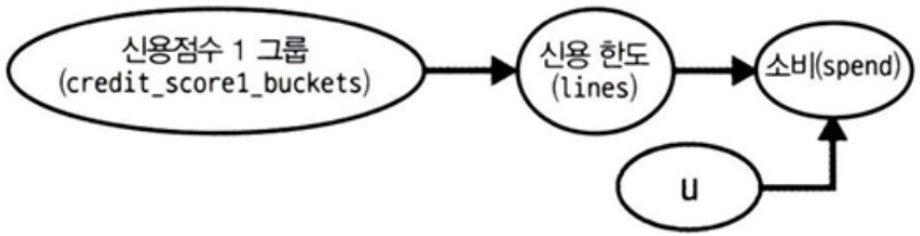
→ 선형회귀분석에서 "결과를 잘 예측할 수 있고 선택편향을 유발하지 않는 변수"를 포함시켜 요인 보정뿐 아니라 잡음을 제거하는 데에도 사용할 수 있다!

중립 통제변수

Causal-Lab

- 통제 변수는 잡음을 줄일 수도 있지만, 늘릴 수도 있음

- 예시. 교란 요인 credit_score1이 교란 요인이 아니라면?



→ credit_score1은 T의 원인이지만 Y이 원인이 아니기에 이를 보정할 필요 X

중립 통제변수

Causal-Lab

- 단순회귀로 제곱근함수를 적용하여 인과효과를 추정한 결과

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2153.2154	218.600	9.850	0.000	1723.723	2582.708
np.sqrt(credit_limit)	16.2915	2.988	5.452	0.000	10.420	22.163

- (비교) 교란요인이 아닌 credit_score1_buckets을 포함시킨 결과

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2367.4867	556.273	4.256	0.000	1274.528	3460.446
C(credit_score1_buckets)[T.200]	-144.7921	591.613	-0.245	0.807	-1307.185	1017.601
C(credit_score1_buckets)[T.400]	-118.3923	565.364	-0.209	0.834	-1229.211	992.427
C(credit_score1_buckets)[T.600]	-111.5738	570.471	-0.196	0.845	-1232.429	1009.281
C(credit_score1_buckets)[T.800]	-89.7366	574.645	-0.156	0.876	-1218.791	1039.318
C(credit_score1_buckets)[T.1000]	363.8990	608.014	0.599	0.550	-830.720	1558.518
np.sqrt(credit_limit)	14.5953	3.523	4.142	0.000	7.673	21.518

→ 표준오차 증가하여 인과 매개변수의 신뢰구간 넓어짐 (OLS는 처치의 분산이 큰 그룹을 선호하기 때문)

→ 신뢰구간이 넓어진다는 것은 결과의 불확실성이 증가한다는 것

→ 모델의 정확도를 높이기 위해서는 처치 T를 잘 설명하는 교란 요인을 통제하면, 분산을 줄일 수 있음

특성 선택: 편향-분산 트레이드 오프

Causal-Lab

- 현실적으로는 처치에만 영향을 주고 결과에는 영향을 주지 않는 공변량 X를 찾아보기 어려움

- 처치 T를 정확히 추정하려면, 모든 교란 요인을 고려해야 함

- 하지만 교란 요인이 너무 많으면,
  모델의 복잡성이 증가하고 추정값의 표준오차가 커지거나 분산이 증가할 수 있음 주의

요약

Causal-Lab

- 예측 도구가 아닌 교란 요인 보정 및 분산 감소 관점에서의 회귀분석 역할
- 조건부 독립성이 유지될 때, 직교화를 이용해 처치가 무작위로 배정된 것처럼 보이게 할 수 있음

  → T를 X에 회귀하여 편향 제거된 T인 잔차를 구하여 X로 인한 교란편향을 보정

FWL 정리에 따른 다중회귀분석의 3단계

(1) 편향 제거 단계에서 처치 잔차 T(~) 구하기

(2) 잡음 제거 단계에서 결과 잔차 Y(~) 구하기

(3) 결과 모델 단계에서 결과 잔차 Y(~)를 처치 잔차 T(~)에 회귀하여

T가 Y에 미치는 인과효과 추정값 도출