SFERS DATA Seminar with Python: 5

Yitae Kwon

SFERS of SNU

2024-1

Contents

Nonparametric Inference

Parametric Statistics

우리가 일반적으로 접하는 통계학은 '**모수통계학**'(parametric statistics)입니다. 추정과 검정의 과정에서 표본의 분포가 정규분포와 같은 특정 분포를 따름을 가정하고, 그 위에서 결론을 얻어내고는 합니다.

예를 들어, 정규분포 $N(\mu, \sigma^2)$는 두 개의 미지의 **모수**(parameter) μ와 σ^2를 안다면 완벽히 기술할 수 있습니다.

e.g. 정규분포를 따르는 IID 표본 $X_1, \cdots, X_n$에서 평균의 추정량:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Parametrical Statistics

모수(parameter)는 특정 분포의 특성을 설명/요약할 수 있는 모든 값들을 말합니다. 모수는 실수가 될 수도 있고, 가끔은 벡터나 행렬이 되기도 합니다. 일반적으로 모수는 θ처럼 씁니다. **모수모형**(parametric model)은 이러한 모수들 유한 개로써 표현될 수 있는 모형을 의미합니다. **모수공간**(parametric space)는 모수가 속할 수 있는 공간을 의미하며, θ를 포함한다는 의미에서 Θ로 자주 씁니다.예를 들어, 정규분포 $N(\mu, \sigma^2)$의 모수는 μ와 σ^2이며, 모수공간은 $\mathbb{R} \times [0, \infty)$입니다.

어떤 함수가 모수 θ와 상수 α에 의하여 변할 수 있다면, 우리는 해당 함수를

$$f(x_1, \cdots, x_n; \theta, \alpha)$$

처럼 써
(i) f가 θ, α에 의해 함수 그 자체로 변화하면서
(ii) $x_1, \cdots, x_n$이라는 가변적인 값에 의하여 함숫값이 결정됨
을 동시에 표현합니다.

t-test for Two Sample

두 모집단의 평균을 비교하기 위한 **t-test**를 생각하여 봅시다. t-test는 두 모집단이 모두 정규분포를 따른다는 **모수적 가정** 하에서 이루어집니다.

e.g. 경제학부 A반과 B반의 평균 학업성적 차이를 알아보고자 한다.
검정에서 모수적 가정은 아래와 같다.

1. 경제학부 A반의 학업성적 분포는 $N(\mu_1, \sigma_1^2)$이다.
2. 경제학부 B반의 학업성적 분포는 $N(\mu_2, \sigma_2^2)$이다.

귀무가설과 대립가설은

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

으로 주어지며, A반 학생들의 학업성적은 $X_1, \cdots, X_{n_1}$이라는 표본으로, B반 학생들의 학업성적은 $Y_1, \cdots, Y_{n_2}$라는 표본으로 얻어질 수 있다.

t-test for Two Sample

만약 $\sigma_1^2 = \sigma_2^2$임이 알려져 있다면, 검정통계량

$$t = \frac{\bar{Y} - \bar{X}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

이 자유도 $n_1 + n_2 - 2$인 t분포를 따름을 이용하여 검정할 수 있습니다. **pooled estimator** s_p는

$$s_p = \sqrt{\frac{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}{n_1 + n_2 - 2}}$$

인 $\sigma_1^2(\sigma_2^2)$의 가장 효율적인 불편추정량입니다. 동시에, t-test는 모형이 맞다는 가정 하에 **검정력**(power)이 가장 높은 검정입니다.

Evaluating Tests

통계적 검정을 평가할 때에는 두 가지 척도를 들여다봅니다.

- **제1종의 오류율**: 귀무가설이 참일 때 검정이 귀무가설을 기각할 확률
  $\Rightarrow$ **유의수준** α에서 검정을 수행하라는 것은 곧 해당 검정의 제1종의 오류율이 α 이하이도록 설계하라는 것입니다.

- **제2종의 오류율**: 대립가설이 참일 때 귀무가설이 기각되지 않을 확률
  $\Rightarrow$ **검정력**(power)은 실제 모수가 θ일 때 1-(제2종의 오류율)을 의미합니다. 즉, θ의 함수로 주어집니다. 일반적으로 검정력은 표본크기 n이 증가함에 따라 증가합니다.

Evaluating Tests

통계적 검정은 아래의 방식으로 평가합니다.

1. 검정이 제1종의 오류율을 원하는 수준에서 통제하는가?
2. 검정의 검정력은 어떠한가?

만약 첫째 평가 기준을 통과하지 못했더라면, 검정의 **기각역**(critical region)을 바꾸어 유의수준을 통제합니다. 첫째 평가 기준을 통과한 검정들 사이에서는 검정력을 비교합니다. 자주 사용되는 것은 **점근상대효율**(Pitman's asymptotic relative efficiency)입니다. 이는 주어진 대립가설 $\theta = \theta_0$와 주어진 검정력 β에 도달하기 위하여 필요한 표본의 크기 n을 비교하는 방법입니다.

Nonparametric Statistics

t-test는 모수적 가정(정규분포 가정)이 참일 때 검정력이 가장 좋은 검정입니다. 만약 경제학부 학생들의 학업성적 분포가 정규분포가 아니라면 어떻게 될까요? 그 분포의 모양조차 알 수 없는 상황이라면? t-test의 검정력이 약해질 수도 있고, 제1종의 오류율이 통제되지 않을 수도 있습니다.

비모수통계(nonparametric statistics) 혹은 **분포무관 방법**은 그 해법을 제공합니다. 이들은 모수적 가정을 최대한 줄인 채 통계적 추정과 검정을 수행하는 과정을 일컫습니다. 비모수통계학에서의 가정은 매우 약합니다. 정규분포로 분포의 모양을 한정시켰던 t-test와 달리, 비모수적 검정에서는 분포족을

$$\mathcal{F}_1 = \{f : \int x^2 f(x)dx < \infty\} \quad \text{(평균과 분산이 존재하는 모든 분포)}$$

$$\mathcal{F}_2 = \{f : f(\mu + x) = f(\mu - x)\} \quad \text{(중앙값 } \mu\text{를 기준으로 대칭인 분포)}$$

와 같이 매우 넓게 설정할 수 있습니다.

Robustness

표본 $X_1, \cdots, X_n$을 관측했을 때, 그 관측값이 다른 관측값들과는 매우 이질적인 관측값을 **이상점**(outlier)라 부릅니다. 만약 모집단의 분포가 꼬리가 긴 분포라면, 극단값이 나올 확률이 크므로, 이상점이 나올 확률 또한 높습니다. 그러나 꼬리가 얇은 분포인 정규분포에 기반한 많은 검정, 추정들의 경우 이러한 이상점들을 제대로 다루지 못합니다.

반면 비모수적 방법은 이러한 꼬리가 긴 분포에 대해서도 유의수준을 잘 통제하므로, 이상점에도 **강건**(robust)하게 추정과 검정을 수행할 수 있습니다. 대표적으로, 중앙값은 관측값의 일부가 오염(이상점)되더라도 추정값이 크게 변하지 않으므로 평균보다 로버스트합니다.

Recommended Books

아래는 비모수통계학 및 통계적 추정/검정이론 등에서 참고할 수 있는 자료들입니다.

- 모두를 위한 컨벡스 최적화: https://convex-optimization-for-all.github.io/
- 비모수통계학 with R: https://product.kyobobook.co.kr/detail/S000001762578
- An Introduction to Statistical Learning: https://www.statlearning.com/
- The Elements of Statistical Learning: https://hastie.su.domains/Papers/ESLII.pdf
- Bootstrap Methods in Econometrics: https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/5/239/files/2020/04/EC11CH08_Horowitz193-224-2_Corrected.pdf

Bootstrap

우리는 가끔 통계량의 표본분포를 구해야 할 때가 있습니다. 예를 들어 정규분포가 가정된 모집단에서 표본을 뽑아 그 표본평균을 보는 상황에서, 이 표본평균으로써 모평균의 95% 신뢰구간을 구하려면 표본평균이라는 통계량의 표본분포를 알아야만 합니다. 모수적 가정이 있는 상태라면 표본평균이 정규분포를 따르며 평균이 모평균과 같고 분산은 모분산을 표본의 크기로 나눈 값임을 알지만, 비모수통계에서는 이것이 불가능합니다. 애초에 해당 표본평균의 실제 분포가 유한 개의 모수로 표현되지 않을 수도 있습니다. 비모수적으로 표본평균의 분포를 근사하는 방법이 바로 **붓스트랩**(Bootstrap)입니다.

Bootstrap

Bootstrap의 기본 가정부터 먼저 세팅해 보도록 하겠습니다.

- $X_1, X_2, \cdots, X_n \sim_{i.i.d} F$
- $\hat{\theta}_n = T(\mathsf{X}_n) = T(X_1, \cdots, X_n)$은 표본으로부터 얻은 모수 θ의 추정량

즉 모집단 F에서 크기가 n인 표본 X_n을 뽑아냈고 그것을 이용해 추정량 $\hat{\theta}_n$을 구한 상황입니다. 그런데 우리는 대부분의 상황에서 F는 모르기 때문에, 수리통계적인 계산을 통해 표본분포를 구하기가 어렵습니다. 혹은 F가 너무 복잡해서 표본분포를 계산, 혹은 기술하기 어려운 상황입니다. 그렇다면 이러한 표본을 여러 개 새로 뽑아 $(\mathsf{X}_n^{(1)}, \mathsf{X}_n^{(2)}, \cdots, \mathsf{X}_n^{(B)}$처럼 크기가 n인 표본을 B개) 몬테카를로 방법으로써 표본분포를 근사할 수도 없는 노릇입니다. 표본을 새로 뽑는 일이 가능할지도 의문이고, 비용도 많이 드니까요.

Bootstrap

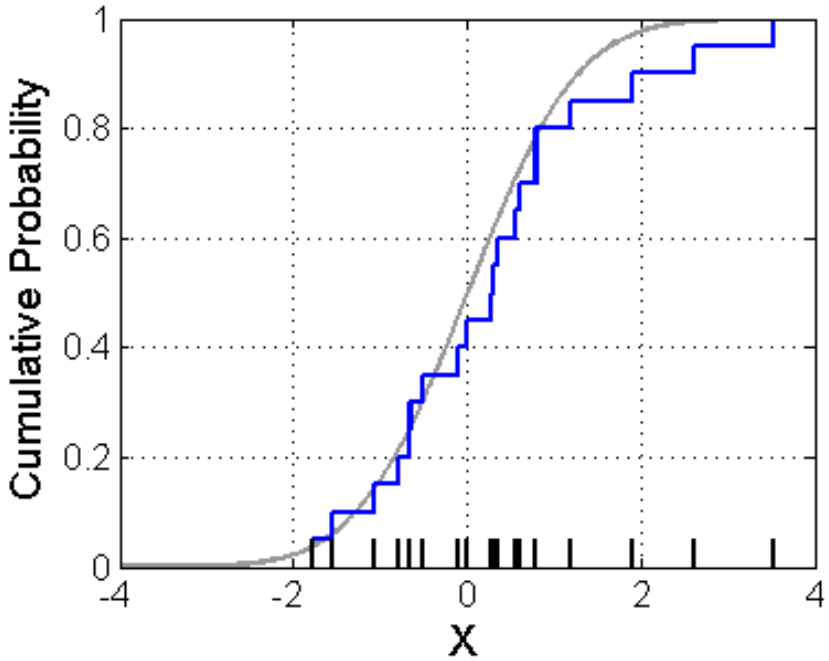
여기서 Bootstrap의 아이디어는, 우리가 F를 모르기 때문에 F에서 표본을 뽑는 게 불가능하니까, 표본 X_n으로부터 모집단의 분포 F를 추정하여 그 추정된 분포 F_n을 가상의 모집단으로 두고 여기서 B개의 크기가 n인 표본을 추출해보자는 것입니다. **경험적 분포함수**

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

는 표본 X_n에서 추정할 수 있는 $F(x)$의 좋은 추정량입니다. 일반적으로 둘의 차이가 대략 $1/\sqrt{n}$임이 알려져 있습니다. 즉 표본의 크기가 충분히 큰 상황에서는 F에서 뽑으나, F_n에서 뽑으나 사실 별 차이가 없다는 의미입니다.

Bootstrap

F_n을 조금 더 잘 들여다 보겠습니다. 이는 사실 x의 값이 특정 X_i를 지날 때마다 $1/n$씩 증가하는 계단형 함수입니다. 이러한 형태의 CDF를 가지는 함수는 바로 이산분포입니다. 즉 F_n은 관측값 X_i에 질량 $1/n$이 배정된 이산분포와 같습니다.



Bootstrap

우리는 이제 이 이산분포를 기반으로, X_n에서 서로 독립인 n개의 관측값을 단순복원추출하여 **붓스트랩 표본** $X_n^{*(i)}$를 얻어내고, 이로부터 통계량 $\hat{\theta}_n^{*(i)}$를 $i = 1, 2, \cdots, B$에 대해 계산합니다. 굳이 n개의 관측값을 뽑는 이유는, 표본분포 역시도 n개의 관측값으로부터 뽑힌 표본으로 얻은 통계량들이 만드는 분포이기 때문입니다. 이제 우리는 B개의 $\hat{\theta}_n^{*(i)}$로부터 얻어낸 경험분포를 $\hat{\theta}_n$의 **붓스트랩 표본분포**라 부르며, $\hat{\theta}_n$의 표본분포의 추정값으로 사용할 것입니다. 이때에도 경험분포를 사용해 추정하는 과정에서 오차 $1/\sqrt{B}$가 생성됩니다. 우리는 이를 G^*라고 부르겠습니다. (G는 원래 추정량 $\hat{\theta}_n$의 분포)

Bootstrap

$$\text{Population} : F$$

$$\Downarrow \text{ Choose } n \text{ observations to build sample with sample size } n \Downarrow$$

$$\text{Sample} : \mathbf{X}_n = (x_1, x_2, \cdots, x_n)$$

$$\Downarrow \text{ Under } F_n \text{ instead of } F, \text{ make B bootstrap samples with sample size } n \Downarrow$$

$$\text{Error} : \frac{1}{\sqrt{n}}$$

$$\text{Bootstrap sample } 1 : \mathbf{X}_n^{*(1)} = (x_1^{*(1)}, \cdots, x_1^{*(n)}), \quad \hat{\theta_n}^{*(1)} = T(\mathbf{X}_n^{*(1)})$$

$$\vdots$$

$$\text{Bootstrap sample } B : \mathbf{X}_n^{*(B)} = (x_1^{*(B)}, \cdots, x_1^{*(B)}), \quad \hat{\theta_n}^{*(B)} = T(\mathbf{X}_n^{*(B)})$$

Bootstrap

$$\text{Bootstrap sample } 1 : \mathbf{X}_n^{*(1)} = (x_1^{*(1)}, \cdots, x_1^{*(n)}), \quad \hat{\theta_n}^{*(1)} = T(\mathbf{X}_n^{*(1)})$$

$$\vdots$$

$$\text{Bootstrap sample } B : \mathbf{X}_n^{*(B)} = (x_1^{*(B)}, \cdots, x_1^{*(B)}), \quad \hat{\theta_n}^{*(B)} = T(\mathbf{X}_n^{*(B)})$$

$$\Downarrow \text{ Estimate boostrap sample distribution using } (\hat{\theta_n}^{*(1)}, \cdots, \hat{\theta_n}^{*(B)}) \Downarrow$$

$$\text{Error} : \frac{1}{\sqrt{B}}$$

$$\text{Estimate: } G^*(x) = \frac{1}{B} \sum_{i=1}^{B} I(\hat{\theta_n}^{*(i)} \leq x)$$

Estimation with Bootstrap

붓스트랩의 가장 기본적인 아이디어는

$$\text{모집단} \Leftrightarrow \text{표본}$$

과

$$\text{표본} \Leftrightarrow \text{붓스트랩 표본}$$

사이의 상동성을 들여다 보는 것입니다.

Estimation with Bootstrap

먼저 표본분포의 평균 $\mu_{\hat{\theta}_n}$은 아래와 같이 추정할 수 있겠습니다. 즉 B개의 붓스트랩 표본에서 얻은 B개의 추정량 $\hat{\theta}_n^*$들의 표본평균, 혹은 적률추정량, 혹은 몬테카를로 추정량을 구한다고 생각할 수 있습니다.

$$\hat{\mu}_{\hat{\theta}_n} = \mu_{\hat{\theta}_n^*} = \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_n^{*(i)} = \int \hat{\theta}_n^* dG^*$$

Estimation with Bootstrap

다음으로는 편향 bias($\hat{\theta}_n$)을 추정해 보고자 합니다. 조심해야 할 것은, 여기서 구하는 편향은 붓스트랩을 시행하기 전, 크기가 n인 표본 X_n에서 얻은 $\hat{\theta}_n$의 편향입니다. 편향은 만약 우리가 F를 알았다면 그냥 계산을 통해 구할 수 있었겠지만, 지금의 우리는 F도 모르고 표본을 더 뽑을 수도 없는 상황입니다. 사실 추정하려는 모수 θ도 모르는 상황입니다. 따라서 편향 계산에 사용 표본분포의 기댓값과 모수 θ를 사용할 수 없습니다.

$$\begin{aligned}
\text{bias}(\hat{\theta}_n) &= E(\hat{\theta}_n) - \theta \\
&= \mu_{\hat{\theta}_n} - \theta \\
&\approx \mu_{\hat{\theta}_n^*} - \hat{\theta}_n \quad \text{(상동성!!)} \\
&= \mu_{\hat{\theta}_n^*} - \hat{\theta}_n = \widehat{\text{bias}}_{\text{Boot}}(\hat{\theta}_n)
\end{aligned}$$

따라서 위에서처럼 $\mu_{\hat{\theta}_n}$의 좋은 추정량이 $\mu_{\hat{\theta}_n}{}^*$이고, θ의 좋은 추정량이 $\hat{\theta}_n$임을 이용하여 편향을 추정하게 됩니다.

Estimation with Bootstrap

한편 이제는 표준오차를 추정해 보도록 하겠습니다. 이 경우 $(\hat{\theta}_n^{*(1)}, \cdots, \hat{\theta}_n^{*(B)})$를 G^*로부터 비롯된 크기가 B인 표본이라고 보고 그 표본표준편차를 구하여 이를 표준오차의 추정량으로 이용합니다. B가 크다면 G^*이 G와 충분히 유사하기 때문에, $\hat{\theta}_n^*$들을 G로부터 비롯된 B인 표본이라 보고 구하면 근사적으로 G의 표준오차를 추정할 수 있게 되는 것입니다. 따라서

$$\widehat{\mathrm{se}}_{\mathrm{Boot}}(\hat{\theta}_n) = \left(\frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_n^{*(i)} - \mu_{\hat{\theta}_n^*})^2 \right)^{\frac{1}{2}} \left(\approx \left(\int (\hat{\theta}_n - \mu_{\hat{\theta}_n})^2 dG \right)^{\frac{1}{2}} \right)$$

입니다. 이처럼 붓스트랩을 이용하면 표본분포를 근사할 수 있는 것에서 그치지 않고, 근사한 표본분포를 이용해 원래 추정량인 $\hat{\theta}_n$의 편향과 표준오차를 파악할 수 있습니다.

Bootstrap Confidence Interval

붓스트랩은 점추정량만을 구하는 데 국한되지 않습니다. 붓스트랩으로는 추정량의 신뢰구간도 구할 수 있고, 다양한 방식이 개발되어 있습니다.

정규근사 붓스트랩 신뢰구간(normal-approximated bootstrap confidence interval)은 표본의 크기 n이 충분히 크거나 분포가 대칭이라 중심극한정리에 의한 정규분포로의 수렴이 잘 일어난 상황에서 사용할 수 있습니다. 즉, 추정량의 편향 $\text{bias}(\hat{\theta}_n)$에 대하여

$$\frac{\hat{\theta}_n - \text{bias}(\hat{\theta}_n) - \theta}{\text{se}(\hat{\theta}_n)} \dot{\sim} N(0,1)$$

가정이 잘 성립하는 상황을 고려합니다.

Bootstrap Confidence Interval

그렇다면 기존 통계학에서 사용하던 것과 비슷하게, $1-\alpha$ 신뢰구간을

$$\theta \in \hat{\theta}_n - \text{bias}(\hat{\theta}_n) \pm z_{\alpha/2}\text{se}(\hat{\theta}_n)$$

로 구성할 수 있습니다. 한편 우리는 일반적인 경우 편향과 표준오차를 알지 못하므로, 이를 붓스트랩으로 구하자는 것입니다. 따라서 이를 붓스트랩 추정량으로 대체한

$$\theta \in \hat{\theta}_n - \widehat{\text{bias}}_{\text{Boot}}(\hat{\theta}_n) \pm z_{\alpha/2}\widehat{\text{se}}_{\text{Boot}}(\hat{\theta}_n)$$

가 정규근사 붓스트랩 신뢰구간입니다. 한편 이는 정규근사 가정 하에서 이루어진 것이기에, 표본의 수가 적어 정규근사가 불가능하거나, 추정량의 점근분포가 정규분포가 되지 않는 경우에는 부정확합니다. 일반적으로 이 경우 포함확률은 $1-\alpha+O(n^{-1/2})$ 수준임이 알려져 있습니다. 이런 상황을 **일차 정확도**를 가진다고 표현합니다.

Bootstrap Confidence Interval

다른 방법으로 신뢰구간을 구할 수도 있습니다.

- 표준 붓스트랩 신뢰구간: pivot H의 분포를 이용합니다.
- 붓스트랩-t 신뢰구간: 스튜던트화된 피벗이 있다고 가정하고 그 분포를 이용합니다. 이차 정확도를 가짐이 알려져 있습니다.(포함확률 $1-\alpha+O(n^{-1})$)
- 편향 조정 및 가속 신뢰구간(bias-corrected and accelerated confidence interval; BCa CI): 적당히 편향과 왜도를 조정하여 피벗(이차 정확도)
- 붓스트랩 퍼센타일 신뢰구간(Bootstrap percentile confidence interval)

Bootstrap Confidence Interval

붓스트랩 퍼센타일 신뢰구간은 매우 간단합니다. 모수 θ의 추정량 $\hat{\theta}_n$과 붓스트랩 표본분포 G^*을 붓스트랩으로써 구했다면, 신뢰수준 $1-\alpha$에서 이는 G^*의 분위수함수 $Q_{G^*}(p)$에 대하여

$$\theta \in \left(Q_{G^*}(\frac{\alpha}{2}), Q_{G^*}(1-\frac{\alpha}{2}) \right)$$

로써 주어집니다. 증명은 생략하겠습니다. 이는 적당한 가정 하에서 일차 정확도를 가짐이 알려져 있습니다. 다만 이는 가장 간단하게 구할 수 있는 신뢰구간 중 하나이므로, 가장 많이 사용되고 있는 붓스트랩 신뢰구간 중 하나입니다.

Hypothesis Testing with Boostrap

가설을 검정하는 방법 중 하나는, 가설검정의 유의수준을 α라 할 때 모수에 대한 $1-\alpha$ 신뢰구간을 만든 후 귀무가설 하에서의 모수가 이 신뢰구간에 들어가는지 보는 것입니다. 이는 통계적으로 문제가 없습니다. 따라서 우리는 **신뢰구간을 뒤집음으로써** 가설검정을 수행할 수 있습니다. 즉 근사적으로 포함확률이 $100(1-\alpha)\%$인 신뢰구간이

$$\theta \in (L(G^*), U(G^*))$$

의 형태로 주어진다면, $H_0 : \theta = \theta_0$의 유의수준 α에서의 근사적 검정은

$$\text{reject if } \theta_0 \leq L(G^*) \text{ or } \theta_0 \geq U(G^*)$$

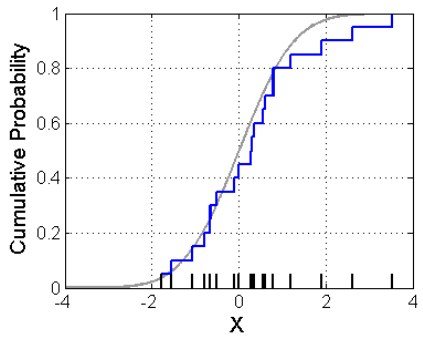
입니다.

Take-Home Messages

- 붓스트랩 관련해서는 정말 간단히만 소개드렸습니다. 궁금하신 내용이 있으시면 위키...(https://en.wikipedia.org/wiki/Bootstrapping_(statistics)) 읽어보세요.
- 붓스트랩은 간단하면서도 쉽게, 추정량의 분포를 모르는 상황에서도 점근적으로 옳은 통계적 추론을 할 수 있다는 장점을 가지고 있습니다. 만약 분포 가정을 하고 싶지 않거나, 데이터가 모양이 이상한 경우 유용하게 이용됩니다.
- 반면 어찌하였든 $B = 1000$번 정도 다시 표집을 하므로, Bn번 수준의 표본추출이 더 필요합니다. 계산량이 엄청나게 증가합니다. 더불어 사실 $n^{-1/2}$나 $B^{-1/2}$ 정도면 그리 만족스러운 수렴 결과가 나오기는 어려운 차수입니다. 다르게 말하면 n이 작으면 애초에 붓스트랩을 사용하기 어렵습니다. 이에 따라 수렴 속도를 빠르게 하기 위한 다양한 해결법들이 제시되고 있습니다.

Kernel Density Estimation

한편 붓스트랩과 몬테카를로 시뮬레이션에서 우리는 그 eCDF를 이용하여 통계량의 표본분포/모수를 근사하는 방법을 배웠습니다. 하지만 eCDF는 불연속한 함수이며, 이에 따라 매끈한 확률밀도함수나 누적분포함수가 필요할 경우 충분하지 않습니다. 특히 eCDF를 pdf로 번역한다면, 이는 단지 거의 모든 점에서 0이고 몇몇 점에서 무한대의 피크가 찍히는 형태가 등장할 것입니다. 이를 위하여 확률밀도함수, 혹은 density function을 비모수적으로 추정하는 방법이 여러 가지 개발되어 있습니다. 오늘은 이 중 가장 잘 알려진 **kernel density estimation**(KDE)에 대해 알아 보겠습니다.



Kernel Density Estimation

자연스럽게 생각할 수 있는 방법 중 하나는 피크의 빈도를 보아 피크가 자주 있으면 확률밀도함수의 값이 크고, 피크가 드물게 있으면 그 값이 작다고 취급하는 것입니다. 이를 다르게 표현하면, 확률밀도함수가 매끈하다고 가정할 때, 아주 작은 값 Δ_x에 대하여

$$f(x - \Delta_x) \approx f(x) \approx f(x + \Delta_x)$$

이므로 $f(x)$의 추정에 $X_i \in (x - \Delta_x, x + \Delta_x)$인 X_i 관측값들을 이용하자는 것입니다. 이처럼 **커널**(kernel)은 주어진 x 주변의 관측값을 반영하는 함수입니다. 주어진 커널 $K(\cdot)$과 주어진 **너비**(bandwidth) h에 대하여,

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

가 커널 밀도함수 추정량입니다.

Kernel Density Estimation

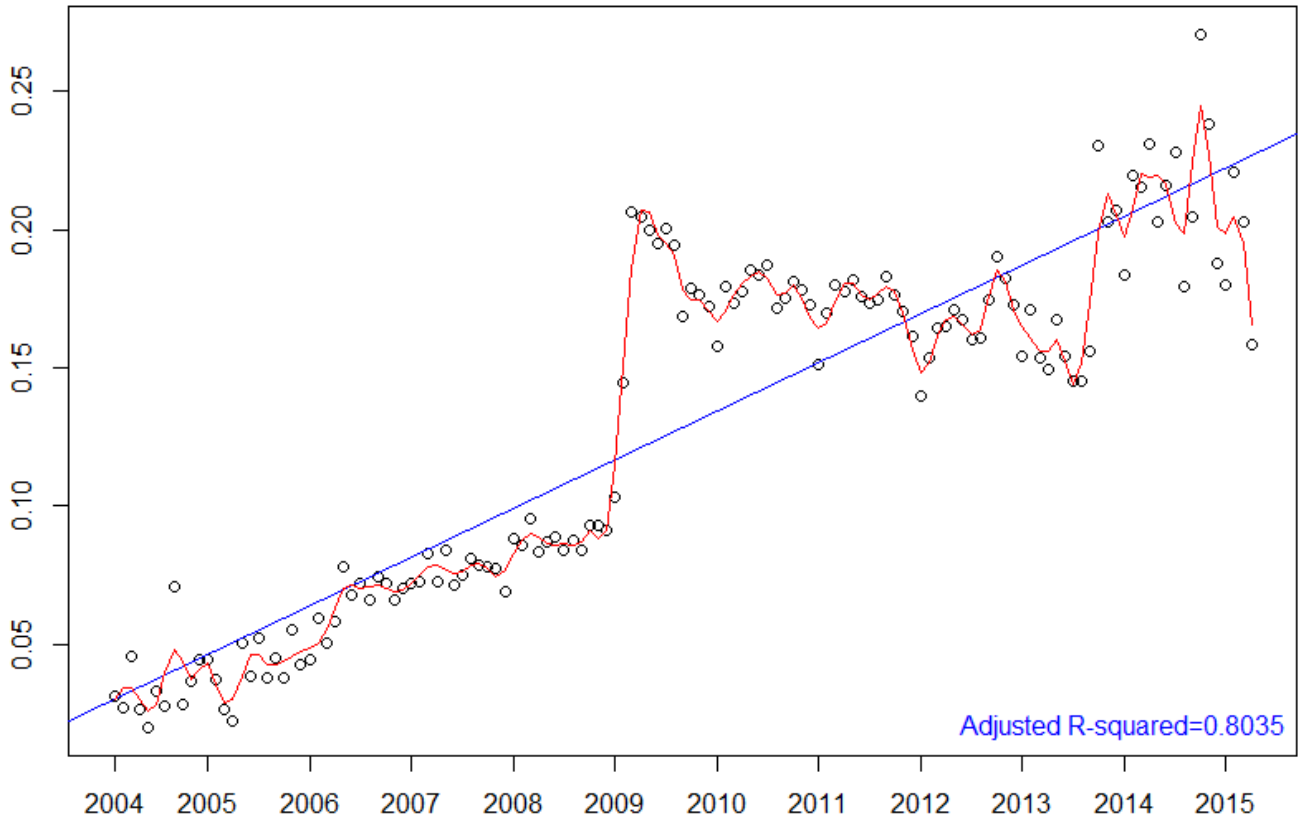
커널은 아래 조건을 만족하는 $K : (-\infty, \infty) \to (0, \infty)$를 커널함수로 이용합니다.

1. $K(x) \geq 0, \int K(x)dx = 1$
2. $\lim_{|x| \to \infty} K(x) = 0$
3. $K(x) = K(-x)$

대표적으로 사용되는 커널은 위 세 조건을 만족하는 **가우스 커널**(Gaussian Kernel)입니다. 이는 그 이름에 걸맞게 표준정규분포를 커널함수의 모양으로 삼고 있습니다. 즉, $K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$입니다. 일반적으로 커널의 형태보다는 너비 h가 그 추정량의 효율에 더 지대한 영향을 미침이 알려져 있어, K보다는 h의 결정을 더욱 중시합니다. h가 커질수록, x에서 먼 X_i들도 유용하게 사용하므로 평활화 수준이 커집니다.

Nonparametric Regression

이를 더욱 확장하면 밀도함수만이 아니라 일반적인 함수들에 대해서도 매끄럽게 추정해줄 수 있습니다.



비모수적 회귀분석

일반적인 회귀분석 식의 형태는 아래와 같습니다.

$$Y = AX + \epsilon$$

그러나 이러한 회귀분석은 선형성 가정의 위반에 로버스트하지 않다는 문제점을 가지고 있습니다. 비모수적 회귀분석에서 추정하고자 하는 것은

$$Y = m(X) + \epsilon, \quad E[\epsilon] = 0$$

에서 $m(\cdot)$이라는 미지의 함수입니다. 이때 m은 선형연산자만이 아니라 모든 매끄러운 함수들을 포함하는 집합 $\mathcal{H}$의 원소입니다. 단순히 말하면, 우리가 산점도를 딱 볼 때 떠올릴 수 있는, 자연스럽고 매끄러운 곡선으로 smoothing을 하겠다는 것입니다.

커널 추정량

자료 (x_i, y_i)를 n개 수집하였을 때, m의 **커널 회귀곡선 추정량**은

$$\hat{m}(x) = \sum_{i=1}^{n} \frac{K_h(x - x_i)}{\sum_{j=1}^{n} K_h(x - x_j)} y_i = \sum_{i=1}^{n} w_i^0(x) y_i$$

처럼 y_i의 가중평균으로 주어집니다. 이때 $K_h(x)$는

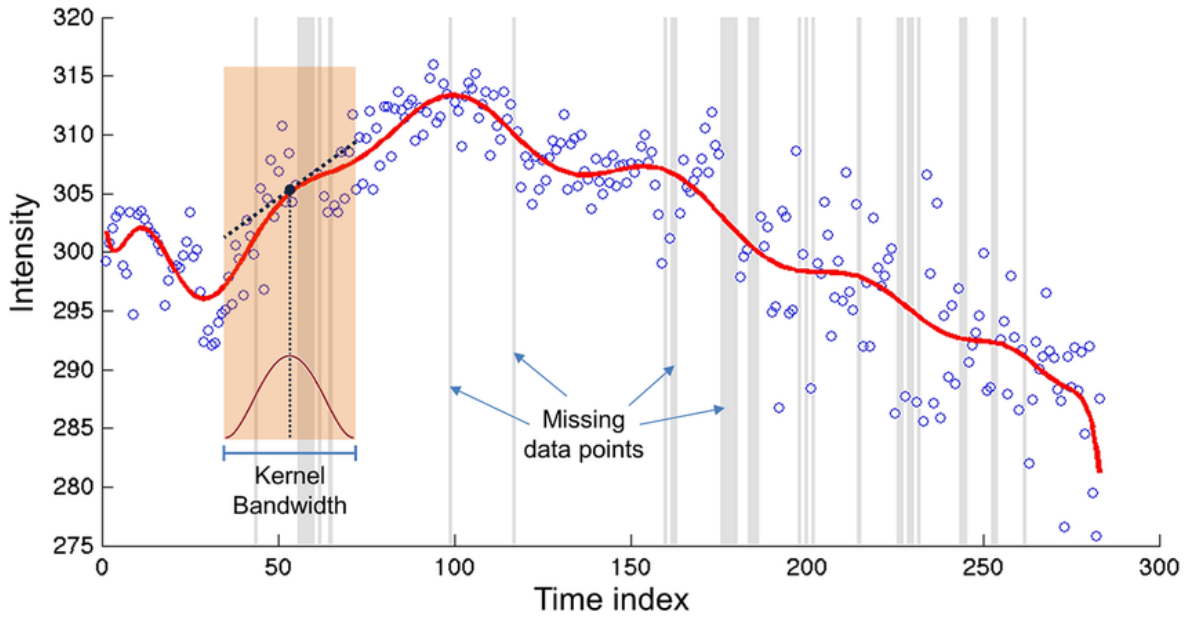
$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$$

입니다. 즉 각 (x_i, y_i)는 x에서의 함숫값을 추정할 때 $K_h(x - x_i)$에 비례하는 가중치를 받습니다. 이를 **나다라야-왓슨 추정량**, 혹은 **국소 상수 추정량**이라 부르기도 합니다. 이를 국소 상수 추정량이라 부르기도 하는 이유는, 각 x 근처에서 국소적으로 볼 때 m이 거의 상수함수라고 가정하고 모형 $y_i = \alpha_x + \epsilon_i$ 하에서 α_x를 구해낸 결과와도 같기 때문입니다.

국소선형회귀 추정량

국소 상수 추정량을 확장하여 국소적으로 선형모형 $y_i = \alpha_x + \beta_x x + \epsilon_i$를 가정할 수도 있습니다. 이 경우 우리의 문제는 아래의 $\hat{\alpha}_x, \hat{\beta}_x$를 찾는 것으로 바뀝니다.

$$(\hat{\alpha}_x, \hat{\beta}_x) = \text{argmin}_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 K_h(x - x_i)$$



국소다항회귀 추정량

마찬가지로 국소적으로 다항회귀를 수행할 수도 있습니다. 그렇게 얻은 추정량을 **국소다항회귀 추정량**이라 부릅니다. 이는 회귀분석에서의 WLS 문제와 동일하며, 쉽게 풀어줄 수 있습니다.

일반적으로 차수가 증가함에 따라 편향은 감소하는 반면, 분산은 증가하는 trade-off가 존재합니다. 다만 좋은 조건이 만족될 때 홀수인 차수에서의 분산과 하나 낮은 짝수인 차수에서의 분산이 같습니다. 따라서 국소선형회귀 추정량, 국소3차회귀 추정량과 같은 홀수차수 추정량이 자주 선호됩니다.

국소다항회귀 추정량

국소다항회귀 추정량에서의 MSE는 차수만이 아니라 K의 형태, support, h의 선택, 그리고 x값 등에 따라 천차만별입니다. 이때 x값의 선택에 의한 영향을 줄이고, 전역에서의 평가를 위해 **MISE(Mean Integrated Square Error)**를 자주 이용합니다.

$$MISE(\hat{m}) = \int_{\mathcal{X}} MSE(\hat{m}(x))dx = \int_{\mathcal{X}} E[(\hat{m}(x) - m(x))^2]dx$$

국소선형회귀 추정량의 MISE는

$$\text{MISE}(\hat{m}_1|\mathbf{X}_n) \approx h^4 \int \frac{(m''(x))^2}{4} f(x)dx + \frac{\int \sigma^2(x)dx}{2\sqrt{\pi}} \frac{1}{nh}$$

로 주어집니다. 이때 f는 X_i의 pdf입니다.

국소다항회귀 추정량

h로 미분하면 적절한 커널의 너비를 찾을 수 있습니다.

$$h^* = \left(\frac{\int \sigma^2(x)dx}{2\sqrt{\pi} \int (m''(x))^2 f(x)dx} \cdot \frac{1}{n} \right)^{\frac{1}{5}}$$

나머지 항들은 그냥 적절히 추정해서 넣는(plug-in) 경우가 대부분입니다.
중요한 것은 $h^* \propto n^{-1/5}$라는 것입니다.

다른 추정량들

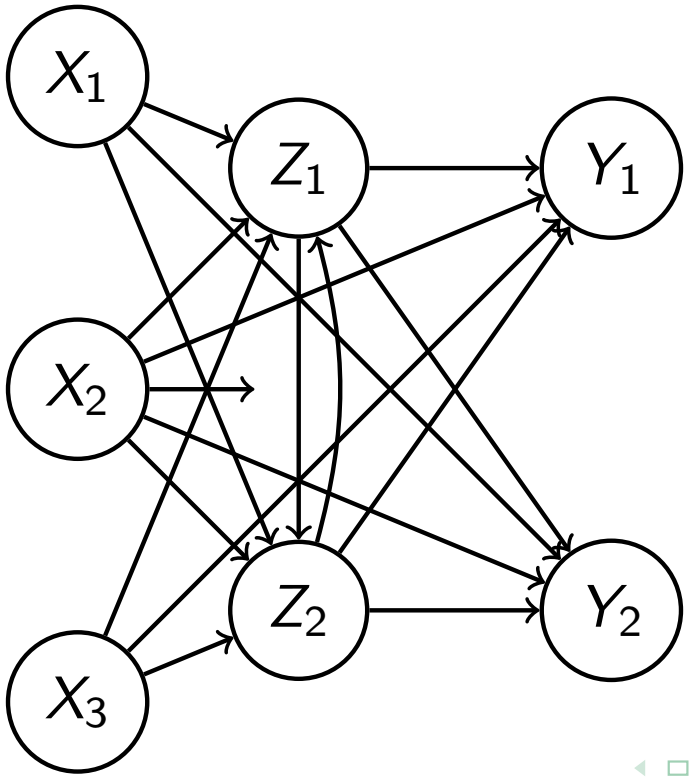
- multivariate인 경우 이를 확장하여 수행하면 됩니다. 하지만 차원의 저주 때문에 추정량에 문제가 생기는 것은 매한가지입니다. 최근에는 자료 구조에 어떠한 다양체를 주어서 내재적 차원을 추정하고, 그 다양체 위에서 차원에 맞게 regression을 수행하는 방법도 연구되고 있습니다.
- Kalman filtering을 smoothing에 적용할 수 있습니다.
- 부분부분별로 나누어 spline method를 적용할 수 있습니다.
- wavelet을 이용할 수 있습니다.
- additive model, partially linear model 등... 주의할 점은 m이 어떤 형태냐에 따라 linear regression과 같은 특정 모형이 좋을 수도 있고, 그렇지 않을 수도 있다는 것입니다. 단, '로버스트'함의 관점에서는 확실히 비모수적인 추정량들이 우위에 섭니다.

Python Practice

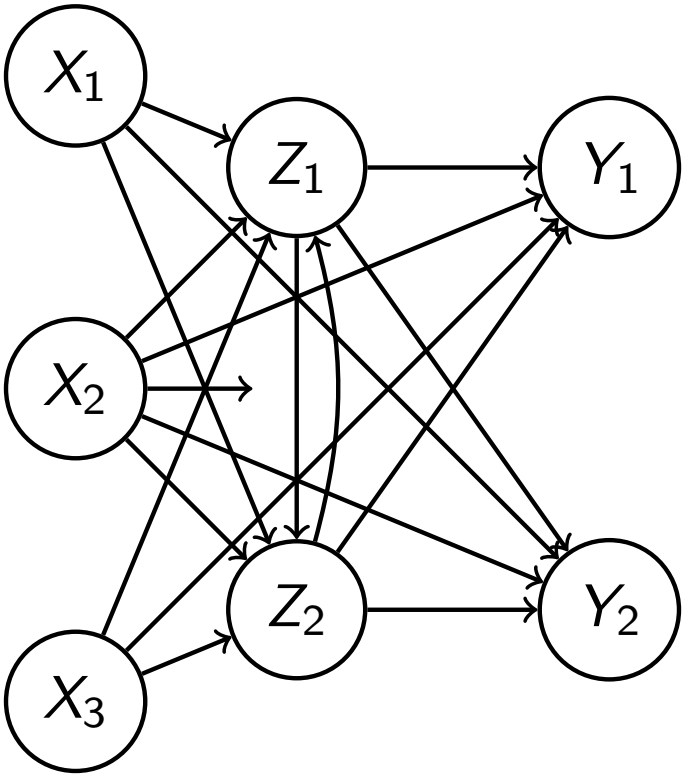
Causal Inference on Complex DAG

Structural Equation Model

우리가 DAG를 그려본다고 하여도, 항상 다중회귀분석이나 도구변수를 통한 분석이 가능한 것은 아닙니다. 특히 가격과 거래량의 관계처럼 서로가 서로에게 영향을 미칠 수 있다거나, 처리 효과가 다른 변수를 통한다든가, 처리의 효과가 해당 피험자의 성질에 의해 달라질 수도 있씁니다. 이러한 복잡한 DAG가 있을 때 그 분석을 위해 사용되는 도구 중 하나가 **구조방정식**(structural equation model)입니다.



Structural Equation Model



이전과 다른 점에는 무엇이 있을까요?

- Z_1와 Z_2가 상호 간에 영향을 미칠 수 있습니다. (동시성)
- 이에 따라 Z_1은 Y_1에 직접효과를 가지는 것과 동시에, Z_2를 통한 간접효과도 가집니다. (**매개효과**; mediation)
- X_2는 Z_1의 Z_2에의 영향을 바꿀 수 있습니다. (**조절효과**; moderation)

Structural Equation Model

Elements

- y: endogeneous variables (모형 내에서 다른 요소의 변화에 의하여 변할 수 있는 요소. Z도 사실 여기에 포함됨)
- X: exogeneous variables (모형 외부에서 주어지는 요소.)
- u: 오차항

이제 목표는 어떠한 Y_i에 u에 의한 **shock**가 발생하거나, X_i가 특정 수준만큼 변했을 때 특정 Y_j에 미치는 영향을 분석하는 것으로 변합니다.

여기에서 관점에 따라 조금씩 모델링이 달라집니다. 첫째는 시계열 자료하에서 이러한 분석을 수행하는 것이고, 둘째는 일반적인 횡단면 자료에서 모형화하는 것입니다. 여기에서는 횡단면 모형에서 SEM을 **simultaneous equation model**으로 바라보고 적합하는 방식을 논의해 보겠습니다. 시계열에서의 일반적인 SEM은 다음 시간에 SVAR/SVECM에서 논의하겠습니다.

Structural Equation Model

그렇다면 우리는 m개의 y, n개의 X가 있다고 가정할 때 아래의 가장 간단한 회귀모형을 만들어줄 수 있습니다.

$$
\begin{aligned}
y_1 &= \alpha_1 + \beta_{1,2} y_2 + \beta_{1,3} y_3 + \cdots + \beta_{1,m} y_m + \gamma_{1,1} X_1 + \cdots + \gamma_{1,n} X_n + u_1 \\
y_2 &= \alpha_2 + \beta_{2,1} y_1 + \beta_{2,3} y_3 + \cdots + \beta_{2,m} y_m + \gamma_{2,1} X_1 + \cdots + \gamma_{2,n} X_n + u_2 \\
&\vdots \\
y_m &= \alpha_m + \beta_{m,1} y_1 + \beta_{m,2} y_2 + \cdots + \beta_{m,m-1} y_{m-1} \\
&\quad + \gamma_{n,1} X_1 + \cdots + \gamma_{n,n} X_n + u_m
\end{aligned}
$$

이때 내생성은 이들이 서로 영향을 주고받으면서 동시에 이 식들을 만족시킬 수 있는 상황에서 평형을 형성함에 따라 발생합니다. 즉 그 경로는 내생변수 y_i들을 통합니다.

Structural Equation Model

이제 우리가 전 페이지의 식을

$$y = \alpha + By + \Gamma X + u$$

와 같이 쓴다면, B는 대각원소가 0인 $m \times m$ 행렬이고, γ는 크게 제약이 없는 $n \times n$ 행렬입니다. 주의할 것은 아직 데이터의 개수 N은 등장하지도 않았다는 것입니다. 이제 내생성의 원인이 되는 우변의 By를 제거하면, 식은

$$(I - B)y = \alpha + \Gamma X + u$$

가 되고 X는 외생변수이기에 더이상 내생성이 없습니다. 따라서

$$y = \underbrace{(I - B)^{-1}\alpha}_{\tilde{\alpha}} + \underbrace{(I - B)^{-1}\Gamma}_{\tilde{\Gamma}} x + \underbrace{(I - B)^{-1}u}_{\tilde{u}}$$

처럼 쓸 수도 있습니다. 앞선 형태를 **structural form**, 뒤의 형태를 **reduced form**이라고 부릅니다.

Structural Equation Model

우리는 causal diagram을 그린 뒤, 이에 맞게 α, B, Γ에 적절한 형태로 constraint를 가해야 합니다. 특정한 변수가 다른 변수에 영향을 주지 않을 것으로 기대된다면, 특정 β나 γ가 0이라는 제약을 걸고 추정해줄 수 있습니다.

만약

1. cycle이 없는 DAG 형태이면서
2. 내생성이 없도록 u가 서로 독립이라면

이를 **recursive model**이라고 합니다. 이는 내생성 문제가 없으므로 그냥 OLS로 추정해주면 됩니다. 반면 앞선 경우처럼 일반적인 경우에는 reduced form 하에서 분석을 진행해 주어야 합니다.

이제 문제는 우리가 non-recursive model 하에서 $\tilde{\alpha}$, $\tilde{\Gamma}$, $\tilde{u}$을 추정하면 다시 α, Γ, B를 복원할 수 있냐는 것입니다. 우리가 알고 싶은 건 Γ와 B이지, $\tilde{\Gamma}$가 아닙니다.

Structural Equation Model

우리가 추정하고자 하는 변수는 B에서 $m^2 - m$개, Γ에서 n^2개, α에서 m개로, 총 $n^2 + m^2$개입니다. 다르게 말하면, $\tilde{\alpha}, \tilde{\Gamma}, \tilde{u}$에 대한 정보로부터 총 $n^2 + m^2$개의 방정식을 얻어야 합니다.

이제 $\tilde{\alpha}$에서는 m개, $\tilde{\Gamma}$에서 n^2개의 정보를 알 수 있으므로, $\tilde{u}$로부터 $m^2 - m$개의 추가적인 방정식을 얻어야 합니다.

$$\text{Var}(\tilde{u}) = \text{Var}((I - B)^{-1}u) = (I - B)^{-1}\text{Var}(u)(I - B)^{-T}$$

$\text{Var}(u) = I$로 가정한다면, 우리는 $\tilde{u}$에 대하여 관찰되는 공분산 행렬 $\tilde{\Sigma}$가

$$\tilde{\Sigma} = (I - B)^{-1}(I - B)^{-T}$$

이도록 $(I - B)^{-1}$를 결정할 수 있습니다. 이는 총 $\frac{m^2+m}{2}$개의 제약을 제공합니다.(Why?) 즉 제약이 $\frac{m^2}{2} - \frac{3m}{2}$개만큼 부족합니다. 우리는 보통 이를 위해서 causal diagram에서 몇 개의 연결을 의도적으로 끊어 변수가 0이도록 함으로써 이만큼의 제약을 추가합니다.

Structural Equation Model

즉 일반적인 과정은 아래와 같습니다.

1. 모형을 기반으로 하여 DAG, 혹은 causal diagram을 그린다.
2. cycle이 없는 DAG 형태라면, 내생변수 간 선후관계가 존재하므로 recursive model 하에서 OLS를 적합한다.
3. cycle이 있다면, causal diagram에서 약할 것으로 생각되는 몇 개의 연결을 끊어 제약을 건 뒤 reduced form에 대해 OLS를 적용한다.
4. 제약을 바탕으로 $\tilde{\alpha}, \tilde{\Gamma}$에 대한 OLS 추정량에서 α, B, Γ를 복원한다.

다만 이는 제약을 바탕으로 회귀분석을 수행하고 identify하는 과정이 OLS와는 달리 매우 복잡합니다. 따라서 이러한 과정을 통한다는 정도만 확인하시면 되겠습니다. 하나 관찰할 수 있는 것은, 만약 B에 제약을 가해 X가 특정 Y에만 영향을 주게 만든다면, 이는 도구변수로 관찰할 수 있다는 것입니다. 즉 도구변수법은 SEM의 일종으로 볼 수 있습니다.

Structural Equation Model

우리가 도구변수를 필요로 했던 이유는 무엇이었나요? confounder X를 관측할 수 없기 때문이었습니다. X가 내생변수이든, 외생변수이든, 우리가 관측하지 못한다면 SEM 역시 말짱 도루묵일 뿐입니다. 우리는 많은 경우 **proxy**와 같은 형태로 다른 변수를 조사하고, 해당 X의 변동을 해당 proxy가 설명해줄 수 있을 것으로 기대하고 모형을 세워 적합합니다. 통계학에서는 이처럼 모형 내외에서 관측되지 않지만 관측되는 변수들에 영향을 미칠 수 있는 변수들을 **잠재변수**(latent variable)이라고 부릅니다.

특히 심리학과 같은 학문에서는 많은 변수들을 수치적으로 계량화하기 어렵습니다. 이에 따라 '공격성', '감수성'과 같은 추상적 성질을 일종의 잠재변수, 혹은 **factor**로 바라보고, 우리가 관측 가능한 변수가 이들 요소의 결합으로 어떻게 설명되는지 확인합니다. 우리는 시간의 문제로 다루지 못했지만 **인자분석**(factor analysis) 역시도 SEM에 사용될 수 있습니다.

Structural Equation Model

SEM에서 가장 많이 사용되는 모형은 latent variable이 포함된 **LISREL**(LInear Structural RELations) 모형입니다.

Setup

- N: 관측값의 개수
- m: latent endogeneous variable의 개수
- n: latent exogeneous variable의 개수
- p: latent endogeneous variable에 의해 만들어지는 observable endogeneous variable의 개수
- q: latent exogeneous variable에 의해 만들어지는 observable exogeneous variable의 개수

Structural Equation Model

Setup

- η: latent endogeneous variables, m차원
- ξ: latent exogeneous variables, n차원
- ζ: structural disturbances
- y: observable endogeneous variables, p차원
- X: observable exogeneous variables, q차원
- ϵ: measurement errors in endogeneous variables
- δ: measurement errors in exogeneous variables

$$\eta = \alpha + B\eta + \Gamma\xi + \zeta$$
$$y = \Lambda_y \eta + \epsilon$$
$$X = \Lambda_x \xi + \delta$$

Structural Equation Model

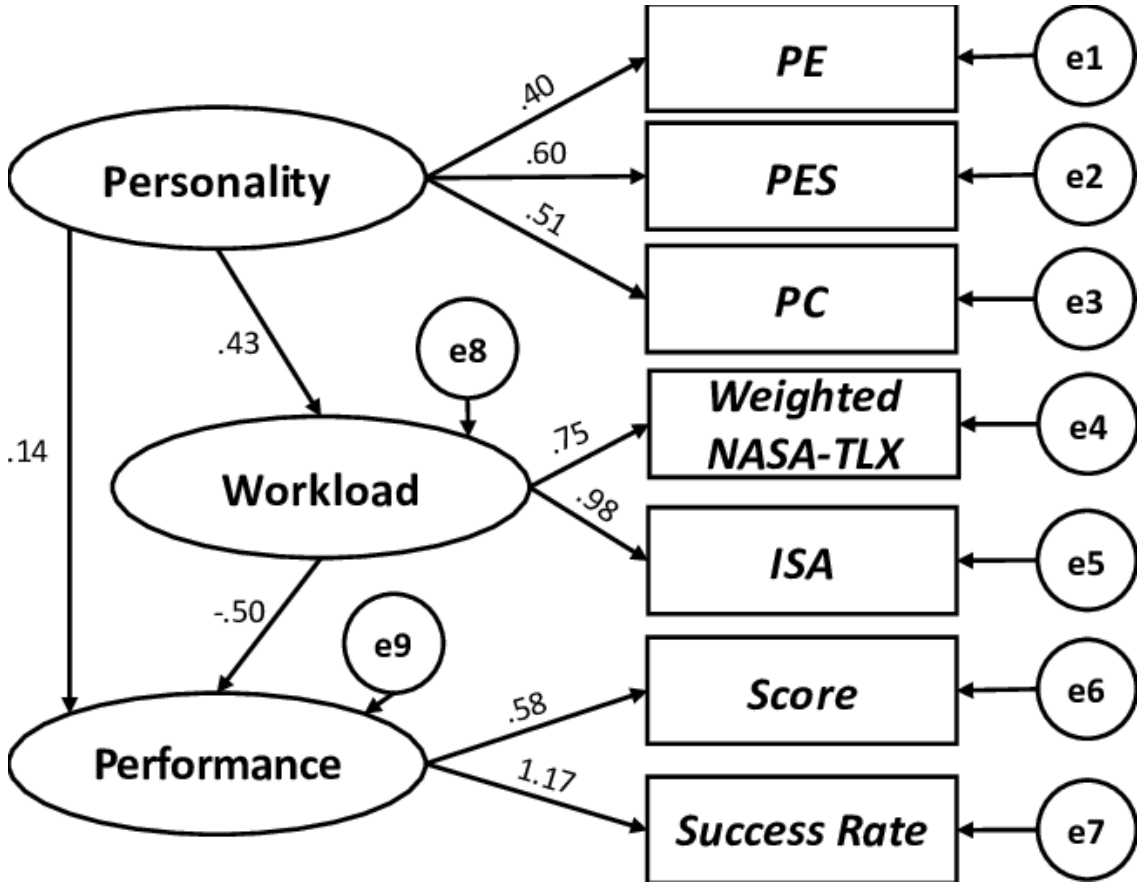
$$\eta = \alpha + B\eta + \Gamma\xi + \zeta$$
$$y = \Lambda_y \eta + \epsilon$$
$$X = \Lambda_x \xi + \delta$$

이때 y와 X는 latent variable들을 우리가 볼 수 있게 해준다는 점에서 **indicator**라 불리기도 합니다. 한편 ϵ과 δ는 이들 **measurment equations**에 의해 y와 X가 결정될 때의 오차라는 점에서 **measurment error**라 불리기도 합니다. Λ_x와 Λ_y는 factor η와 ξ가 X와 y에 미치는 비중을 의미한다는 점에서 **factor loading**이라 불립니다.

잠재변수가 없을 때와 마찬가지로, $(I - B)^{-1}$의 존재성, ζ와 같은 error term의 structure에 대한 적절한 가정과 제약이 주어져야 적합이 가능함은 동일합니다.

Structural Equation Model

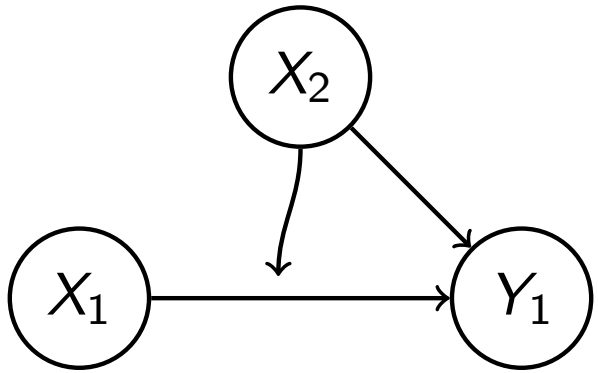
결론적으로 우리는 아래와 같은 구조적 모형을 얻게 됩니다.



Python Practice

Moderation

이제 다시 간단한 다이어그램으로 돌아와 보겠습니다. SEM의 스피릿은 결국 복잡한 모형도 적절한 제약과 함께라면 선형모형으로 모든 분석이 가능해진다는 것입니다. 다시 우리는 엄청나게 간단한 아래의 causal diagram으로 돌아오도록 하겠습니다.



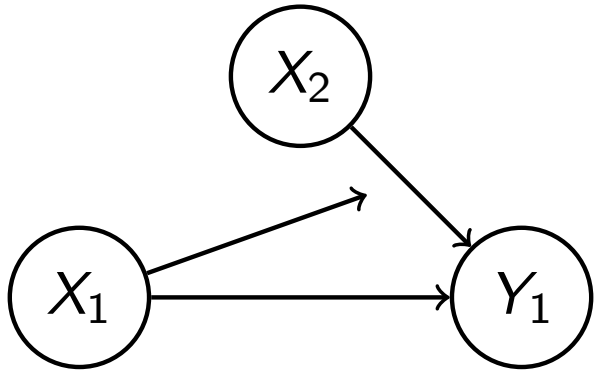
즉 latent variable이 없고, X_1, X_2가 직접 관측 가능한 외생변수이며, Y_1은 관측 가능한 내생변수입니다. 이때 주의할 점은 $X_1 \to Y_1$ 반응경로에 X_2가 포함된다는 것입니다. 이러한 표기는 X_2의 값에 따라 $X_1 \to Y_1$ 효과가 달라질 때를 의미합니다. 이때 X_2를 **moderator**(조절변수)라 부릅니다.

Moderation

이 경우 우리는 아래처럼 선형모형을 만들 수 있습니다.

$$\begin{aligned} Y_1 &= \alpha + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + u \\ &= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u \end{aligned}$$

주의할 것은 이 식은 사실 아래의 causal diagram을 묘사하기도 한다는 것입니다. 일반적으로 연구 목표에 맞게 둘 중 더 적절한 모형을 선택하여 묘사합니다. 만약 둘 중 더미변수가 있다면, 더미변수 쪽을 조절변수로 둡니다.



Moderation

$$Y_1 = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

우리는 단순히 여기에 OLS를 적합하면 $\beta_1, \beta_2, \beta_3$의 일치추정량을 얻을 수 있습니다. 그 해석은 어떻게 할까요? 편의상 X_2가 남자면 1, 여자면 0인 변수라 두겠습니다.

- 남자일 경우의 회귀식: $Y_1 = (\alpha + \beta_2) + (\beta_1 + \beta_3) X_1 + u$
- 여자일 경우의 회귀식: $Y_1 = \alpha + \beta_1 X_1 + u$

조절변수 성별의 Y_1에의 직접효과는 β_2, $X_1 \to Y_1$ 경로에의 **조절효과**는 β_3으로 쓸 수 있습니다. 이들이 유의한지 확인하면 조절효과가 유의한지 판단해줄 수 있습니다.

한편 남자인 경우 X_1의 효과가 유의한지 확인하기 위해서는 $\beta_1 + \beta_3$이 유의한지 보아야 합니다. 이러한 복잡한 검정은 2주차에 다루었으니, 이를 참고해 주세요.

Moderation

한편 이러한 모형들은 기존과 달리 처리의 효과가 개체, 혹은 그 성질에 따라 달라질 수 있음을 의미하고 있습니다. 특정한 공변량 X를 가진 개체들의 처리효과는 **CATE**(conditional average treatment effect)

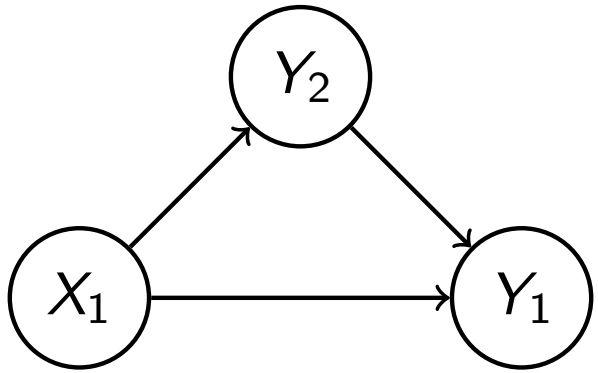
$$CATE(X) := \mathbb{E}[Y(1) - Y(0)|X]$$

으로 정의합니다. 교호작용항을 추가한 moderation effect model은 이러한 CATE를 추정하는 데 도움을 줍니다. CATE가 달라지는, 즉 개체의 성질에 따라 처리 효과가 달라지는 것을 **treatment effect heterogeneity**, 혹은 **effect modification**이라 부르며, 인과추론의 주요 연구 주제 중 하나입니다.

한편 인과추론에서 구분해야 할 것 중 하나는 **treatment**와 **risk factor**입니다. 처리는 우리가 바꿀 수 있는 것, 혹은 잠재변수에 의하여 변할 수 있는 모형에 포함된 것이며, 위험인자는 외생적으로 각 개인에게 결정된(성별 등) 입니다. 일반적으로 위험인자를 조절변수로 넣고, 위험인자에 따라 처리(혹은 exposure)의 효과가 달라지는 것을 보는 것이 국룰입니다.

Mediation

둘째로 moderation과 혼동할 수 있지만 다른 개념인 **mediation**을 알아 보겠습니다. mediation은 한 변수의 다른 변수에의 영향이 제3의 변수를 거쳐 일어나는 상황을 묘사합니다. 이를 causal diagram으로 그리면 아래와 같습니다.



그리고 회귀모형은

$$Y_2 = \alpha_2 + \gamma_2 X_1 + u_2$$
$$Y_1 = \alpha_1 + \beta_1 Y_2 + \gamma_1 X_1 + u_1$$

Mediation

$$
\begin{aligned}
Y_2 &= \alpha_2 + \gamma_2 X_1 + u_2 \\
Y_1 &= \alpha_1 + \beta_1 Y_2 + \gamma_1 X_1 + u_1 \\
&= (\alpha_1 + \alpha_2\beta_1) + (\gamma_1 + \beta_1\gamma_2)X_1 + (\beta_1 u_2 + u_1)
\end{aligned}
$$

그렇다면

- $X_1 \to Y_1$ 경로의 회귀계수는 γ_1: **직접효과**
- $X_1 \to Y_2 \to Y_1$ 경로의 회귀계수는 $\beta_1 \times \gamma_2$: **간접효과**
- 전체 인과효과는 $\gamma_1 + \beta_1\gamma_2$: **총효과**

입니다. 여기 포함된 $\gamma_1, \beta_1, \gamma_2$가 모두 유의해야지만 해당 DAG가 유효하다고 판정합니다.

Python Practice

Causal Inference in Panel Data

Quasi-Experiment

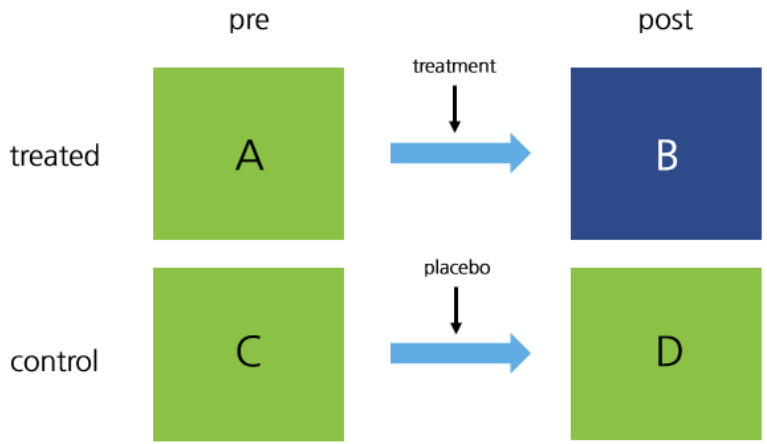
이처럼 아무리 복잡한 causal diagram을 그려서 분석하려고 해도, 내생성을 완벽히 피해가기는 어렵습니다. 우리가 놓치고 있는 부분이 있을 수도 있고, 변수가 제대로 측정되었는지에 대한 의문도 있기 때문입니다. 더불어 논의가 시계열 자료로까지 확장되면, 시간의 흐름에 따른 변화와 처리효과를 구분하기에도 어렵습니다. 따라서 이를 위하여 여러 **quasi-experiment** 세팅이 개발되었습니다.

quasi-experiment는 자연적으로 처리가 주어지는 관찰연구, 혹은 자연실험의 일종으로, RCT는 아니지만 적절한 방식으로 내생성을 통제해 인과추론을 가능하게 하는 실험 디자인을 의미합니다.

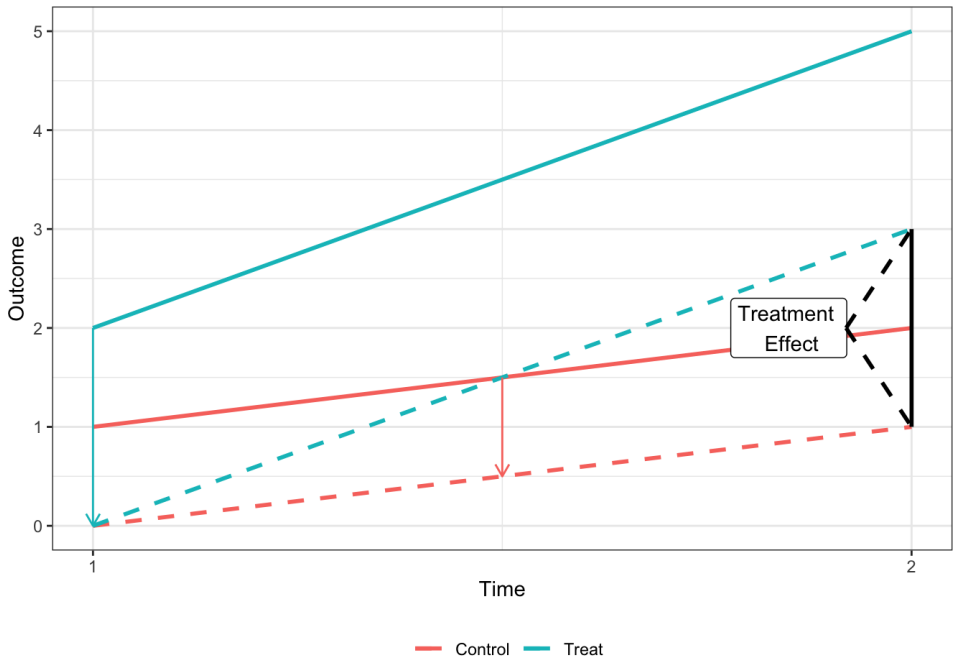
Difference-in-Differences (DiD)

Difference-in-Differences(DiD; 이중차분법)은 가장 많이 사용되는 준실험 방법 중 하나입니다. 기존에 우리가 횡단면 자료에서 분석했던 것과 달리, 이중차분법을 비롯한 뒤의 방법들은 대개 **panel data**가 있어야 사용할 수 있습니다. 다르게 말하면, 각 개체에 대한 정보가 적어도 여러 시점에 있어야만 논의가 가능합니다.

먼저 아래의 상황을 생각해 보겠습니다.



Difference-in-Differences (DiD)



A와 B의 차이는 시간의 흐름에 따른 변화와 처리의 효과가 합해진 것으로 생각할 수 있습니다. 반면 C와 D의 차이는 시간의 흐름에 따른 변화만으로 볼 수 있습니다. 이에 따라, A와 B의 difference와 C와 D의 difference의 차, 즉 difference-in-differences가 처리효과를 의미한다는 것이 기본적인 아이디어입니다.

Difference-in-Differences (DiD)

이를 potential outcome으로 설명해 보겠습니다. 여기에서는 potential outcome에 시간 요소를 넣어 확장해야 합니다. 편의상 pre를 $t = 1$, post를 $t = 2$로 쓰면, potential outcome은 $Y_{i,t}(Z)$처럼 쓸 수 있습니다. 우리가 원하는 개체별 처리효과는

$$Y_{i,2}(1) - Y_{i,2}(0)$$

이 될 것이며, 추정하고자 하는 값은 전향적으로 이 B가 처리를 받음으로써 얻은 평균 처리 효과

$$ATT = \mathbb{E}[Y_2(1) - Y_2(0)|Z = 1]$$

이 됩니다.

Difference-in-Differences (DiD)

DiD에서 가장 중요한 가정은 **parallel trend assumption**(평행 트렌드 가정) 입니다. 이는 만약 treated group이 처리를 받지 않았더라면, control과 mean level에는 차이가 있더라도 유사한 트렌드를 보였을 것이라는 가정입니다. 이를 식으로 쓰면,

$$\mathbb{E}[Y_2(0) - Y_1(0)|Z = 1] = \mathbb{E}[Y_2(0) - Y_1(0)|Z = 0]$$

으로 주어집니다. 이는 앞선 세팅 하에서 시간에 의한 효과가 **가법적**(additive) 하게 주어질 것이라는 가정과 동일합니다.

더불어 숨겨져 있는 가정 중 하나는, 1기에는 아무도 직접적인 처리를 받지 않았기에,

$$Y_{i,1}(1) = Y_{i,0}(0)$$

이 성립합니다. 이를 **consistency** 가정 혹은 **no anticipation** 등으로 부릅니다.

Difference-in-Differences (DiD)

만약 평행 트렌드 가정이 성립한다면,

$$\begin{aligned} ATT &= \mathbb{E}[Y_2(1) - Y_2(0)|Z=1] \\ &= \mathbb{E}[Y_2(1)|Z=1] - \mathbb{E}[Y_2(0) - Y_1(0)|Z=1] - \mathbb{E}[Y_1(0)|Z=1] \\ &= \mathbb{E}[Y_2|Z=1] - \mathbb{E}[Y_2(0) - Y_1(0)|Z=0] - \mathbb{E}[Y_1(1)|Z=1] \\ &= (\mathbb{E}[Y_2|Z=1] - \mathbb{E}[Y_1|Z=1]) - (\mathbb{E}[Y_2|Z=0] - \mathbb{E}[Y_1|Z=0]) \end{aligned}$$

으로 ATT는 difference-in-differences로 identify됩니다. 사실 DiD에서 ATT가 estimand인 이유 역시 이러한 직관적인 추정량과 동일하기 때문입니다. 그러므로 canonical DiD 추정량은

$$\widehat{ATT} = \frac{1}{N_1}\sum_{i=1}^{N} Z_i(Y_{i,2} - Y_{i,1}) - \frac{1}{N_0}\sum_{i=1}^{N}(1-Z_i)(Y_{i,2} - Y_{i,1})$$

으로 주어집니다.

Difference-in-Differences (DiD)

기간이 더욱 여러 개로 증가하고, 처리의 시점이 여러 개로 달라지는 경우에도 이러한 DiD 디자인을 사용할 수 있습니다. 특히 최근에는 이러한 복잡한 세팅 하에서의 DiD 디자인이 많이 개발되었고, 핫한 주제 중 하나입니다.

다만 우리는 시간의 부족으로 간단한 경우만 고려해 보겠습니다. 바로 또또 회귀분석을 이용하는 방법입니다. DiD 세팅 하에서 outcome $Y_{i,t}$는 처리 여부 Z_i 뿐만이 아니라, i번째 개체가 가지고 있는 특성들, 혹은 t번째 시간에 나타난 특정 현상들에 의해 지배될 수 있습니다. 만약 처리 효과가 모두 동질하고, outcome이 i, t, Z_i로 완전히 묘사될 수 있다면,

$$Y_{i,t} = \alpha_i + \alpha_t + \beta D_{i,t} + \epsilon_{i,t}$$

라는 회귀분석을 돌려 얻은 $\hat{\beta}$는 $\widehat{ATT}$와 동일합니다. 주의할 것은 이때 $D_{i,t}$는 실제로 처리를 받았는지 여부로, $Z_i = 1$이더라도 $t = 1$이면 $D_{i,t} = 0$입니다. $t = 2$에서 비로소 $D_{i,t} = 1$이 됩니다.

Difference-in-Differences (DiD)

이 회귀분석식은 **TWFE**(Two-Way Fixed Effects) model이라 불리기도 합니다. 왜냐하면, 개체의 영향을 α_i로 고정된 효과로, 시간의 영향 α_t을 고정된 효과로 바라보고 내생성의 통제를 위해 회귀분석식에 추가하였고, 이에 따라 two-way로 내생성을 통제하는 fixed effect model이 되었기 때문입니다.

이는 parallel trend assumption 역시 보장합니다. 왜일까요? 시간의 흐름에 따른 효과 α_t가 개체의 특성이나 처리 여부와는 무관하게 α_t로써 더해지기 때문입니다. 만약 $D_{i,t} = 0$이라면, 각 개체는 베이스라인이 α_i이고 α_t에 의해서만 이동하게 됩니다.

Difference-in-Differences (DiD)

앞선 모형은 처리 이후 β만큼이 상승하는 모형을 묘사하는 **static TWFE** 입니다. 반면 처리 이후 처리가 기폭제가 되어 outcome의 지속적인 증가가 발생할 수 있습니다. 이러한 경우에는 처리 이후의 시간인 **event-time**에 따라 다른 처리 효과를 부과하는 **dynamic TWFE** 모형을 사용할 수 있습니다.

$$Y_{i,t} = \alpha_i + \alpha_t + \sum_{e=0}^{L} \beta_e I(E_{i,t} = e) + \epsilon_{i,t}$$

이때 $E_{i,t}$는 i번째 개체가 시점 t에서 처리를 받은 지 얼마나 지났는지를 의마하며, β_e는 처리 e시점 후의 처리 효과를 의미합니다.

이외에도, 처리 효과에 존재하는 잠재적인 effect modification을 통제하기 위하여 공변량 X 등을 여기에 추가하는 방법 등도 존재합니다. 역시 스피릿은 내생성을 통제하는 모형이라면 어떤 모형이든 사용할 수 있다는 것입니다.

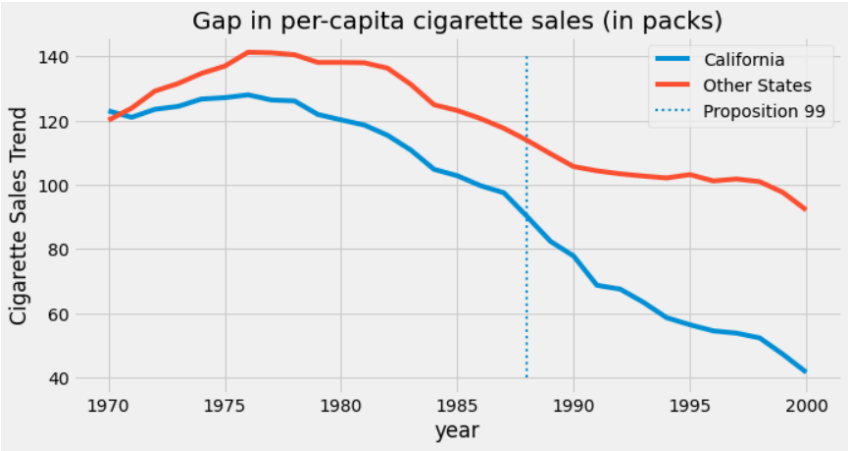
Python Practice

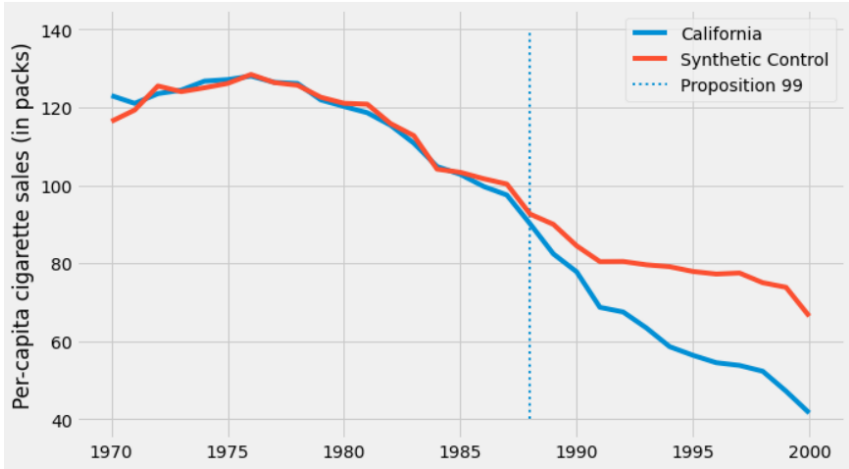
Synthetic Control Method

DiD 디자인은 내생성에 대한 책임을 모두 α_i와 α_t에 넘겨 버리기 때문에, 간단하게 사용할 수 있어 인기가 많습니다. 그러나 가장 큰 문제 중 하나는 평행 트렌드 가정이 만족하지 못하는 데이터가 많다는 것입니다. 이러한 경우에는 counterfatual 부분을 추정할 수 있는 근거가 부족합니다. 다른 데이터를 바탕으로 이 부분을 메꾸자는 것이 바로 **synthetic control**의 아이디어입니다.

DiD에서 평행 트렌드 가정이 만족한다는 것은 무슨 말일까요? 곧 두 개체가 처리군이든 대조군이든, 평균 레벨을 제외하면 시간에 따른 영향 등을 거의 동질하게 받는다는 것입니다. 우리는 이제 다른 데이터들을 이용하여, 처리군과 거의 동질한 반응을 보일 것으로 예상되는 가짜 개체인 **synthetic control unit**을 만들고, 해당 개체가 처리를 받지 않았을 때 어떻게 움직이는지를 처리군의 counterfactual 정보로 사용하려 합니다.

Synthetic Control Method





Synthetic Control Method

Setup

- $J+1$ units in period $t = 1, 2, \cdots, T$
- 첫째 unit은 treatment를 받았고, $T_0 + 1, \cdots, T$에서 처리를 체험
- 나머지 J개의 unit은 **donor pool**으로 작용하는 controls

우리가 원하는 unit의 처리 효과는 $t = T_0 + 1, \cdots, T$에 대하여

$$\tau_{1,t} = Y_{1,t}(1) - Y_{1,t}(0)$$

입니다. 이때 $Y_{1,t}(1)$은 우리가 관측한 값입니다. 따라서 우리는 donor pool의 원소들로써

$$\hat{Y}_{1,t}(0)$$

을 추정하고 이를 바탕으로 $\hat{\tau}_{1,t}$을 얻는 것이 목표가 됩니다.

Synthetic Control Method

우리가 가장 쉽게 생각할 수 있는 방법은 donor pool에 있는 원소들의 선형 결합으로써 이를 만드는 것입니다. 즉

$$W = (w_2, w_3, \cdots, w_{J+1})^T \in \mathbb{R}^J$$

을 $w_j \geq 0, w_2 + \cdots + w_{J+1} = 2$이도록 만든 뒤

$$\hat{Y}_{1,t}(0) = W^T Y_{i,t}(0)$$

을 사용하는 것입니다. 이때 $Y_{i,t}(0)$는 $J \times (T - T_0)$ 차원이 될 것입니다. 이는 곧 donor pool에 있는 원소들의 convex combination을 이용하겠다는 것이나 마찬가집니다. 그러면 W를 어떻게 정하면 좋을까요?

Synthetic Control Method

우리가 DiD를 포기하게 된 이유는 parallel trend가 만족하지 않아서였습니다. 다르게 말하면, W를 parallel trend를 만족시킬 수 있게 만들어주면 됩니다. 따라서 첫째 unit의 처리 전 outcome이나 covairate들 k개를 모아 만든 벡터 X_1를, J개 donor group에 대해 동일하게 수행하게 만든 $k \times j$ 행렬 X_0으로써 가장 잘 예측하게 만들어줄 수 있는 W를 선택합니다. 간단히 쓰면,

$$X_1 = \begin{pmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,T_0} \end{pmatrix}, \quad X_0 = \begin{pmatrix} Y_{2,1} & Y_{3,1} & \cdots & Y_{J+1,1} \\ Y_{2,2} & Y_{3,2} & \cdots & Y_{J+1,2} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{2,T_0} & Y_{3,T_0} & \cdots & Y_{J+1,T_0} \end{pmatrix}$$

일 때

Synthetic Control Method

$$\begin{cases} \text{minimize} & \|X_1 - X_0 W\| \\ \text{subject to} & W \geq 0,\ 1^T W = 1 \end{cases}$$

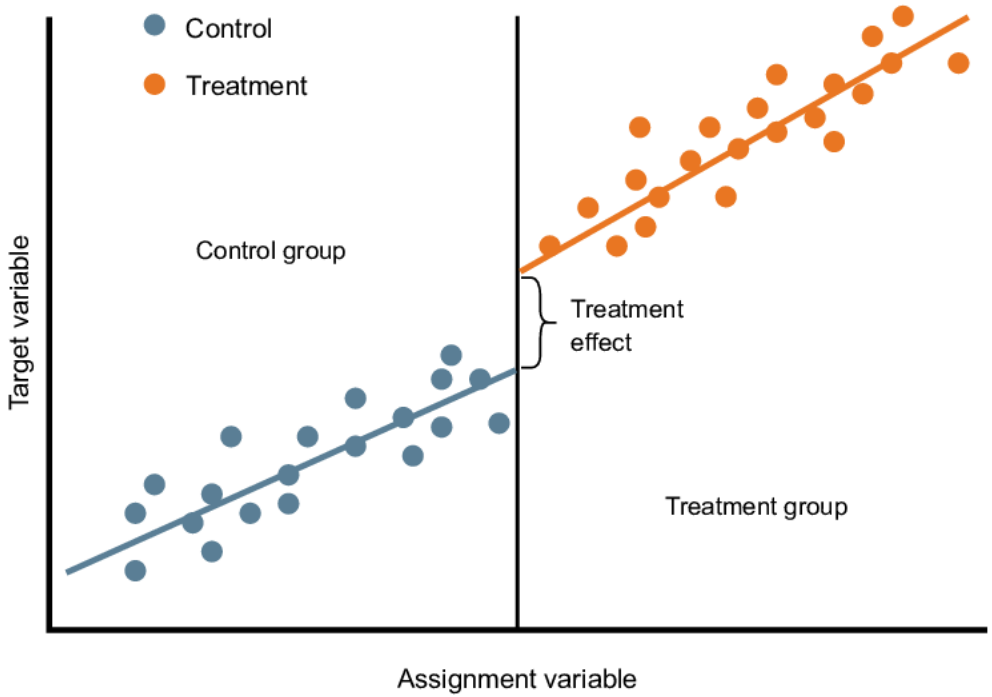
을 만족하는 W를 사용할 수 있습니다. 이는 놀랍게도... 3주차에 배웠던 quadratic programming 혹은 다른 programming 방법으로 풀 수 있습니다. 따라서 컴퓨터로, 혹은 손으로 쉽게 풀 수 있습니다.

설명을 위해 단순히 pre-treatment outcomes들을 넣었지만, 실제로는 각 개인의 변하지 않는 외생변수 등을 추가하여 가짜 대조군을 만드는 데 사용해 줄 수 있습니다.

W에 대한 제약을 없애면 그냥 회귀분석이 되기 때문에 더 쉽다고 생각할 수 있습니다. 그렇지만 이 경우 extrapolation의 문제가 빈번하게 발생합니다. 따라서 donor pool에 존재하는 개체들이 만드는 convex hull에 이 가짜 대조군이 들어오게 하여 추정량의 안정성을 보장하고는 합니다.

Python Practice

Regression Discontinuity Design



내생성은 treatment assignment가 완전히 랜덤하게 이루어지지 않음에 따라 발생할 수 있습니다. 우리는 여기에서 처리 여부가 **assignment variable**의 값에 의해 결정되는 경우를 상상해 보겠습니다.

Regression Discontinuity Design

만약 처리가 assignment variable의 특정 점을 기준으로 이루어진다고 해 봅시다. 그렇다면 우리는

$$Z_i = \begin{cases} 1 & \text{if } X_i \geq x_0 \\ 0 & \text{if } X_i < x_0 \end{cases}$$

처럼 쓸 수 있습니다. 이때 X_i는 assignment variable입니다. 그렇다면 아주 작은 수 ϵ에 대하여, $X_i = x_0 - \epsilon$일 때에는 처리를 안 받고, $X_i = x_0 + \epsilon$일 때에는 처리를 받습니다. 그러나 둘의 차이는 거의 미미하므로, X_i에 의한 outcome의 효과는 없습니다. 따라서 두 경우 outcome의 차이는 오직 Z_i의 추가적 부여로부터 이루어진 것이라고 생각해줄 수 있습니다.

이처럼 특정 **threshold**, 혹은 **cutoff** x_0을 기점으로 처리 여부가 확정적으로 갈리는 디자인을 **sharp RD**라 부릅니다.

Regression Discontinuity Design

그렇다면 이 경우 우리는 x_0을 기준으로 각각 회귀분석을 하여, X_i가 x_0으로 왼쪽/오른쪽에서 감에 따른 값을 구하고, 이를 바탕으로 ATE를 구할 수 있습니다. 사실은 이러한 점 때문에, 우리의 estimand는

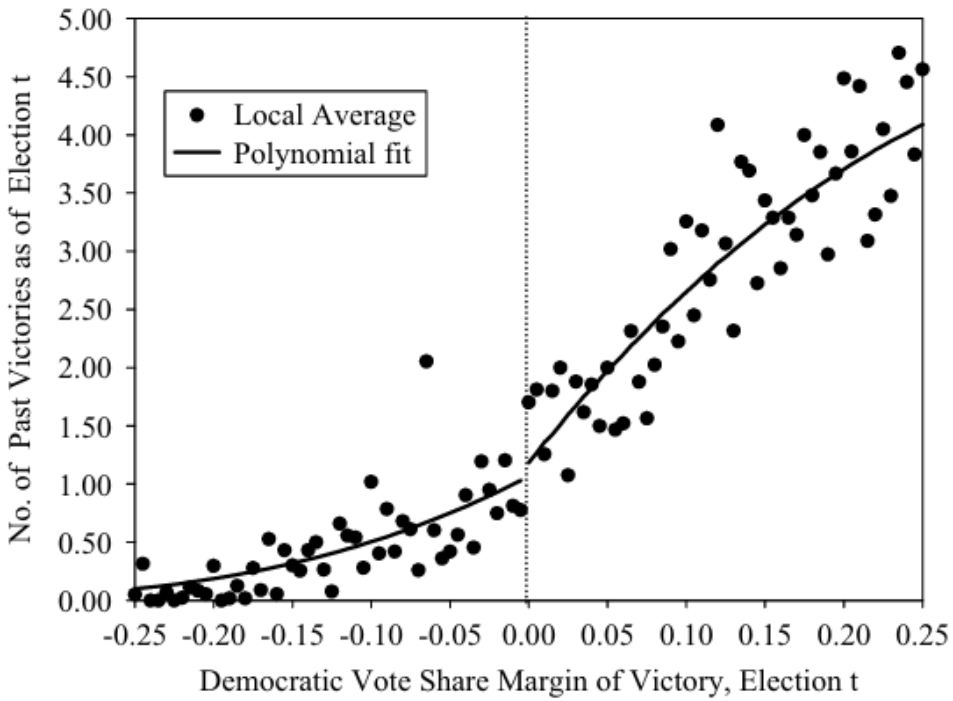
$$CATE(x_0) = \mathbb{E}[Y(1) - Y(0)|X_i = x_0]$$

이 됩니다. 그러나 만약 처리 효과가 모든 점에서 동일하다고 가정한다면, 이는 단순한 ATE와 같고, 그냥

$$Y_i = \alpha + \beta Z_i + \gamma X_i + \epsilon_i$$

라는 회귀모형을 적합하여 얻은 β가 ATE 추정량이 됩니다. 그렇지 않다면, 앞서 배운 nonparametric regression 테크닉을 이용하여 보간해준 뒤 discontinuity가 생기는 지점에서의 CATE를 추정하면 됩니다.

Regression Discontinuity Design



여기에서는 컷오프 전에는 국소 상수 추정량을, 컷오프 후에는 다항식 회귀를 이용하였네요.

Regression Discontinuity Design

한편 Z_i가 X_i에 의해 바로 결정되는 게 아니라, 그를 기점으로 처리 확률이 달라질 수 있습니다. 이 경우에는

$$P(Z_i = 1|X_i) = \begin{cases} g_1(X_i) & \text{if } X_i \geq x_0 \\ g_0(X_i) & \text{if } X_i < x_0 \end{cases}$$

으로 모형화하고, **fuzzy RD**라 부릅니다. sharp RD, fuzzy RD과 같이 cutoff를 바탕으로 assignment mechanism을 조정하는 디자인을 통틀어 **Regression Discontinuity Design**(RDD)라 부릅니다.

fuzzy RD는 $X_i \geq x_0$인지 여부를 곧 도구변수로 바라볼 수 있습니다. 즉 $X_i \geq x_0$인지가 결정되면 처리 확률이 달라지고, 여기에 흡수되어 outcome에 영향을 미친다는 것입니다. 따라서 fuzzy RD에서는 많은 경우 Wald estimator를 사용합니다.

Python Practice

Panel Data Analysis

Spurious Regression

패널자료의 가장 큰 특징 중 하나는 각 개체의 시간에 따른 변화가 나타난다는 것입니다. 우리가 가진 자료가 단순히 다변량 시계열 자료라고 생각해보겠습니다. 우리가 가장 하기 쉬운 접근은 이 시계열 자료를 그냥 횡단면 자료로 보고 회귀분석을 돌리는 것입니다.

이제 $Y_{1,t}$가 t 시점에서의 Y_1, $Y_{2,t}$가 t 시점에서의 Y_2라고 하겠습니다. Y_1과 Y_2 사이의 인과관계를 찾기 위하여 단순하게

$$Y_{1,t} = \alpha + \beta Y_{2,t} + \epsilon_t$$

를 적합하면 될까요? 아닙니다. 주의할 것은 $Y_{1,t}$와 $Y_{2,t}$가 모두 시간에도 의존한다는 점입니다. 즉 시간이 confounder로 기능할 수 있습니다. 시계열 분석 하에서 시간이나 시간의 흐름에 따른 잠재변수의 영향으로 내생성이 발생하는 상황을 **spurious regression**이라 부릅니다.

Error Structure and Clustered Data

한편 패널 데이터에서는 횡단면 데이터, 혹은 시계열 데이터에 비해 error $\epsilon_{i,t}$가 복잡한 경우가 많습니다. 이 error가 random walk를 따를 수도 있고, 이전 시기의 error에 영향을 받을 수도 있고, 개체 i마다 error의 전체적인 값이 달라지는 상황도 존재할 수 있습니다.

예를 들어 서울대학교 졸업생들의 시간에 따른 소득을 패널조사 한다고 해보겠습니다. 그렇다면 여기에는 크게 두 단위의 **cluster**가 존재합니다. 첫째는 각 개체들을 묶는 cluster로, 소속 대학(자연과학대학, 사회과학 등)에 따라 error가 독립적이지 않을 수 있습니다. 각 단과대는 커리큘럼을 공유하기에, 해당 단과대 출신 학생들의 오류는 양의 상관관계를 가질 수 있습니다. 둘째 관점의 cluster는 시간입니다. 경제가 좋을 때에는 전반적인 소득이 상승하고, 불황기에는 반대로 이동하므로, 같은 시간 내의 자료들은 양의 상관을 가질 수 있습니다. 이는 OLS가 아닌 GLS와 같은 복잡한 회귀모형의 적합을 요구하게 됩니다. 물론 스트럭쳐를 알면 그냥 GLS 하시면 됩니다. (2주차 참고)

Seemingly Unrelated Regression

패널데이터에서 각 개체, 혹은 각 시간은 상관이 없는 것처럼 보이지만, 앞서 말한 이유들로 인해 상관관계를 가집니다. 이에 따라 개체별 처리 효과의 이질성을 허용하되 시간에 따른 처리 효과의 이질성은 무시하여

$$Y_{i,t} = \alpha_i + \beta_i Z_{i,t} + X_{i,t}\delta_i + \epsilon_{i,t}$$

라는 모형을 세우고, $\epsilon_{i,t}$에 대해서는

$$\mathrm{Cov}(\epsilon_{i,t}, \epsilon_{j,t}) = \sigma_{i,j,t}, \quad \mathrm{Cov}(\epsilon_{i,t}, \epsilon_{i,s}) = 0$$

이도록 제약을 걸어 FGLS 추정량을 구해주면 됩니다. 이는 각 i에 대한 시계열 자료로 구분해서 구하는 것보다 일반적으로 유리합니다. 특히, $\sigma_{i,j,t}$가 클수록 error structure에 대한 정보를 추가 제공하여 효율이 증가합니다.

각 개체에 대한 회귀식은 보기에는 관련이 없는 회귀식처럼 보이나, 실제로는 cluster/time이 error correlation을 만들기에 합쳐서 추정했을 때 더 효율적입니다. 따라서 이를 **SUR**(Seemingly Unrelated Regression)이라 부릅니다.

Fixed Effects Model

이러한 점들을 모두 고려하는 것은 매우 어렵습니다. 따라서 개체에 따른 cluster와 시간에 따른 cluster를 모두 적절히 error term에서 빼내고, error structure를 OLS에 맞는 간단한 구조로 바꾸는 방법이 생각될 수 있습니다. 가장 간단한 방법 중 하나는 **fixed effects model**(고정효과모형; FEM)입니다.

Setup

- t: 시간
- $Y_{i,t}(Z_{i,t})$: potential outcome
- A_i: unobserved covariates, or latent individual features
- $X_{i,t}$: observed covariates(exogeneous)

Fixed Effects Model

고정효과 모형에서 중요한 가정들은 아래와 같습니다.

- $Z_{i,t}$는 모든 공변량 $A_i, X_{i,t}$에 조건부로는 랜덤화된다.

$$\mathbb{E}[Y_{i,t}(0)|A_i, X_{i,t}, t, Z_{i,t}] = \mathbb{E}[Y_{i,t}(0)|A_i, X_{i,t}, t]$$

- A_i에 선형적으로 $Y_{i,t}(0)$이 결정된다.

$$\mathbb{E}[Y_{i,t}(0)|A_i, X_{i,t}, t] = \alpha + \alpha_t + A_i\delta + X_{i,t}\gamma$$

- 인과관계는 가법적이고, 동질하다.

$$\mathbb{E}[Y_{i,t}(1)|A_i, X_{i,t}, t] = \mathbb{E}[Y_{i,t}(0)|A_i, X_{i,t}, t] + \beta$$

Fixed Effects Model

이 가정들과 함께라면,

$$Y_{i,t} = \alpha_i + \alpha_t + \beta Z_{i,t} + X_{i,t}\delta + \epsilon_{i,t}$$

으로 쓸 수 있고,

$$\alpha_i := \alpha + A_i\delta$$

로 정의됩니다. 그러나 우리는 A_i는 딱히 관심에 없으므로, α_i만 추정해줘도 됩니다. 이처럼 i에 의한 cluster와 t에 의한 cluster 효과를 α_i, α_t로 바꾸는 모형을 **two-way fixed effect model**이라고 부릅니다. 즉 DiD는 FEM의 한 종류입니다. 이때 α_t를 **year effect**, α_i를 **individual effect**라고 부릅니다.

Random Effects Model

우리는

$$\alpha_i = \alpha + A_i\delta$$

로 정의하였습니다. finite population의 관점에서는 α_i가 고정이지만, super-population 하에서 우리가 sampling을 했다고 가정하면 A_i 역시도 random이고, α_i 역시 random일 수 있습니다. **random effect model**(랜덤효과모형; REM)에서는 아래 두 가정을 가져 갑니다.

- unobserved A_i is randomly drawn.
- unobserved A_i is independent to all of the $X_{i,t}$ and $Z_{i,t}$.

즉 A_i에는 오직 $X_{i,t}$로 설명이 불가능한 랜덤한 부분이 포함된다고 생각합니다. 그러면 α_i는 다른 요소와 무관하기에 $\epsilon_{i,t}$에 그냥 편입시킬 수 있습니다.

Random Effects Model

따라서 그냥

$$Y_{i,t} = \alpha_t + \beta Z_{i,t} + X_{i,t}\delta + \epsilon_{i,t}$$

를 적합하면 됩니다. 단, GLS를 통하여 $\epsilon_{i,t}$에 존재하는 clusterd error structure를 고려해주는 것이 더욱 정확합니다.

이것으로부터 알 수 있는 결론은, 그냥 맘 편하게 OLS를 사용하면 되는 fixed effects model을 사용하는 것이 우리 수준에서는 편하다는 것입니다. 물론 엄밀하게 통계학적으로는 finite population인지, super-population인지에 따라 어떤 추정량은 consistent이고, unbiased이고 등등을 따져야 하지만, 응용에서는 그냥 적절한 모형을 사용해서 독자에게 잘 전달해주기만 하면 됩니다.

Dynamic Panel Model

FEM은 omitted variable이 모두 시간에 의해 변하지 않는 A_i 등이라고 가정합니다. 그러나 실제로는 해당 상황이 발생하지 않을 수도 있습니다. 하나의 아이디어는, A_i, 혹은 α_i에 대한 정보가 이전 기의 outcome인 $Y_{i,t-h}$에 저장되어 있다고 믿는 것입니다. 즉

$$\mathbb{E}[Y_{i,t}(0)|Y_{i,t-h}, X_{i,t}, Z_{i,t}] = \mathbb{E}[Y_{i,t}(0)|Y_{i,t-h}, X_{i,t}]$$

을 가정하면

$$Y_{i,t} = \alpha + \alpha_t + \theta Y_{i,t-h} + \beta Z_{i,t} + X_{i,t}\delta + \epsilon_{i,t}$$

가 우리의 회귀모형이 됩니다. 여기에서 $Y_{i,t-h}$를 **lagged dependent variable**이라 부릅니다. 만약 둘 중에 무엇을 넣어야 할 지 고민되면, 그냥 둘 다 넣어서

$$Y_{i,t} = \alpha_i + \alpha_t + \theta Y_{i,t-h} + \beta Z_{i,t} + X_{i,t}\delta + \epsilon_{i,t}$$

를 적합하는 경우도 많습니다.

Dynamic Panel Model

그러나 둘을 다 넣으면 identify에 문제가 생깁니다. 단순히 $h = 1$인 경우를 생각해서 차분하면,

$$\Delta Y_{i,t} = \theta \Delta Y_{i,t-1} + \Delta \lambda_t + \eta \Delta Z_{i,t} + \Delta X_{i,t} \delta + \Delta \epsilon_{i,t}$$

인데 여기서 residual $\Delta \epsilon_{i,t}$가 $\Delta Y_{i,t-1}$와 correlation을 가지기 때문에 OLS로 적합이 불가능하고 inconsistent해집니다. 이를 **Nickell's bias**라 부릅니다.

Dynamic Panel Model

dynamic panel data model(DPD model)의 추정 방법 중 하나인 **Arellano-Bond estimator**(AB approach)는 아래 모형을 세웁니다.

$$Y_{i,t} = \theta Y_{i,t-1} + \eta Z_{i,t} + X_{i,t}\delta + v_{i,t}$$
$$v_{i,t} = u_i + \epsilon_{i,t}$$

즉 위의 모형이지만 fixed effect model with lagged dependent variables를 한 번에 쓰고 OLS 추정하는 대신, 이를 two-step으로 바라봅니다. 또한 $\Delta\lambda_t$ 역시도 각 개체와는 관련없으므로 $v_{i,t}$ 측으로 흡수됩니다.

Dynamic Panel Model

어찌하였든 결국 first-difference를 취하여,

$$\Delta Y_{i,t} = \theta \Delta Y_{i,t-1} + \eta \Delta Z_{i,t} + \Delta X_{i,t}\delta + \Delta \epsilon_{i,t}$$

로 돌아왔습니다. 이제 아이디어는 t시점에서는 이미 결정되어 외생적으로 취급 가능한 $1 \sim t-2$기의 자료들을 **도구변수**로 사용하자는 것입니다. 이 도구변수는 이전의 것이므로 $\Delta \epsilon_{i,t}$와는 무관하면서, $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$과는 상관이 있습니다. 따라서 시간이 T만큼이면 총 $(T-1)(T-2)/2$개만큼의 도구변수가 생기고, IV-GMM을 통해 moment condition

$$\mathbb{E}[Y_{i,s}(\Delta Y_{i,t} - \theta \Delta Y_{i,t-1} - \eta \Delta Z_{i,t} - \Delta X_{i,t}\delta)] = 0$$

을 $t = 2, \cdots, T$와 $s = 1, 2, \cdots, t-2$에 대해 품으로써 η를 얻어내줄 수 있습니다.

Python Practice

Recall: Spirits of Causal Inference

1. 인과관계와 상관관계는 다르다.
2. 모든 모형은 틀리지만, 몇몇 모형은 유용하다. causal diagram을 그리고 시작하자.
3. 가장 전달력이 높은 모형 중 하나는 선형 모형이다.
4. 선형 모형의 회귀분석에서 문제가 되는 것은 내생성이다.
5. 내생성은 생략된 변수들에 의해 발생한다. 오차항에 흡수되어 버린 생략된 변수들은 생략되지 않은 변수들과 assignment mechanism 등에 의해 상관관계를 가질 수 있고, 이는 내생성을 촉발한다.
6. 내생성이 문제가 되는 이유는 OLS 추정량이 일치추정량이 아니기 때문이다. 우리의 모형에 맞는 모형을 사용한다는 것은 곧 인과효과에 대한 일치추정량을 얻을 수 있는 모형을 사용한다는 것이다.

Recommended Pages

- A Guide on Data Analysis
  `https://bookdown.org/mike/data_analysis/`
- Causal Inference for The Brave and True `https://matheusfacure.github.io/python-causality-handbook/landing-page.html`
- Mostly Harmless Econometrics
  `https://www.mostlyharmlesseconometrics.com/`
- 오늘 정말 많은 방법을 정말 간단히만 전달드려서, 각 방법에 대한 깊이는 얕습니다. 분석 방법을 사용하고자 하신다면, 꼭 더 공부를 하신 다음에 pre-test나 diagnostics에 대해서도 고려를 하여 진행하시길 바랍니다. 참고로 오늘 과제는 없습니다.

Next Session

주제: Causal Inference(3) and Time Series Analysis

- Matching
  - Propensity Score and Unconfoundedness
  - Propensity Score Matching(PSM) and Other Matching Designs
- Weighting
  - Outcome Regression
  - Inverse Probability Weighting
  - Doubly Robust Estimator
- Univariate Time Series Analysis
  - Autocorrelation and Stationarity, ARIMA
  - ARCH/GARCH Variants
  - Regression with ARIMA errors vs. ARIMAX
  - Quantile Regression
- Multivariate Time Series Analysis
  - Vector Autocorrelation(VAR) and Structural VAR
  - Granger Causality
  - Cointegration and VECM