Chapter 8

Parametric Bootstrapping: We came, we saw, we CLARIFYed

챕터 8입니다. 어느덧 Lv.2.Stat에서 OLS 파트가 거의 마무리되는 것 같습니다. 챕터 1부터 챕터 7까지 내용을 통해 Lv.1.Stat의 내용들을 복습하거나, 조금은 더 깊은 내용들을 추가해 설명해왔습니다. 여기까지 오셨다면 이제는 신뢰구간이란 무엇인지, 그리고 표집분포란 대체 무엇인지에 대해서 더 명확하게 이해하셨을 것이라 기대합니다. 그리고 앞선 챕터 7을 통해 비모수 부트스트랩까지 다루었기 때문에 굉장히 기초적이지만 이른바 시뮬레이션에 대한 내용도 접해보았습니다.

간단히 복습하는 의미에서 비모수 부트스트랩에 대해 이야기해 보겠습니다. 우리는 모집단에 대해서 알고 싶어합니다. 사회과학, 특히 정치학을 예로 든다면 우리는 민주화에 대해 연구할 때, 민주화의 모집단—모든 민주화에 대해 보편적으로 찾아볼 수 있는 특성을 알기를 원합니다. 하지만 과거의 모든 사례들을 우리가 다 안다고 할 수 없고, 앞으로 일어날 모든 일들을 관측할 수는 없습니다. Karl Popper가 말한 검은 백조의 비유(김웅진 2011: 53)처럼, 우리는 모집단을 결코 완벽하게 관측할 수 없습니다. 우리가 여태까지 관측한 모든 백조가 희다고 한들, 세상 어딘가 검은 백조가 존재할 확률이 0이 아닌 것이죠. 따라서 모집단은 관측불가능하고, 얻을 수 없는 것이라고 해도 무방합니다.

하지만 우리는 모집단에 대해 알고 싶습니다. 그래서 통계학에서는 표본을 이용해 모집단에 대해 추론하는 방법을 사용합니다. 우리의 표본이 모집단으로부터 무작위로 추출되어 수집되었을 때, 우리는 표본이 모집단을 대표할 것이라고 기대합니다. 이때 무작위란 모집단으로부터 추출될 확률이 모든 관측치에 걸쳐 동일하다는 것을 의미합니다.

비모수 부트스트랩은 여기서 한 발 더 나아가 만약 우리가 모집단으로부터 무작위로 반복추출하여 구성한 표본을 가지고 있다고 할 때, 그 표본으로부터 또 무작위 반복추출하여 부트스트랩 표본들을 구성할 경우, 그 부트스트랩 표본들이 모집단으로부터 무작위 추출된 것과 다르지 않다는 논리에 바탕을 두고 있습니다. 즉, 비모수 부트스트랩은 무작위화(randomization)에 대한 강력한 믿음을 전제하고 있습니다.

8.0.1 비모수 부트스트랩을 통해 얻는 것 (1)

비모수 부트스트랩을 통해 우리는 표집분포에 대해 어떠한 함수이든 상관없이 그 전체의 분포에 대한 추정값을 얻을 수 있습니다. 예를 들어, $y = \beta_0 + \beta_1 x + \beta - 2z + \beta 3_x z$라는 상호작용항을 포함한 회귀모델이 있다고 합시다. 이때, 우리가 관심있는 것은 과연 변수들 간 상호작용이 존재하는지를 보여주는 $\frac{\partial y}{\partial x} = \beta_1 + \beta_3 z$일 것입니다.

비모수 부트스트랩을 이용하면, 일단 우리는 $g = 1, \cdots, G$번의 재표집 과정을 거쳐 얻어진 부트스트랩 표본들을 이용해 동일한 모델을 적용해 표본 별 추정값을 재추정하면 그만입니다. 따라서 우리는 1번 부트스트랩 표본부터 G번 부트스트랩 표본에 이르기까지 $\hat{\beta_1}^{(1)} + \hat{\beta_3}^{(1)} z, \cdots, \hat{\beta_1}^{(G)} + \hat{\beta_3}^{(G)} z$라는 결과들을 얻게 됩니다. 남은 것은 이렇게 얻은 부트스트랩 표본들의 추정치 벡터를 평균, 표준편차, 히스토그램, 신뢰구간 등의 다양한 방법을 통해 요약하여 보여주는 일입니다.

8.0.2 비모수 부트스트랩을 통해 얻는 것 (2)

이번에는 OLS를 생각해보겠습니다. $\bar{\mathbf{X}}$가 M개의 행과 K개의 열을 가지고 있는 행렬이라고 하겠습니다. 데이터셋을 생각해보시면 편합니다. 이 행렬은 M개의 ID와 K개의 변수를 가진 데이터셋이라는 의미입니다. 우리는 이 행렬을 가지고 개별 관측치들 각각에 대해서 예측변수들의 수준이 변화할 때, 종속변수의 값이 어떻게 변화하는지를 확인할 수 있습니다.

$g = 1, \cdots, G$의 부트스트랩 표본들을 데이터로부터 얻었다고 하겠습니다. 그 부트스트랩 표본들을 가지고 원래의 데이터와 동일한 OLS 모델을 적용할 경우, 우리는 $K \times 1$에 해당하는 벡터: $\hat{\beta}^{(G)}$를 얻게 됩니다. 하나의 부트스트랩 표본에 우리는 M개의 예측값을 얻을 수 있습니다: $\hat{\mathbf{y}}^{(g)} = \bar{\mathbf{X}}\hat{\beta}^{(g)}$.

데이터로부터 얻은 G개의 부트스트랩 표본들을 가지고 우리는 관심을 가지고 있는 예측값 각각에 대한 전체 부트스트랩 분포를 구할 수 있습니다.

- 부트스트랩 표본들로부터 얻은 계수값들을 $\hat{\mathbf{b}} = (\hat{\beta}^{(1)} \cdots \hat{\beta}^{(G)})$라고 하겠습니다.
- 마찬가지로 부트스트랩 표본들로부터 얻은 예측값을 $\hat{\mathbf{Y}} = \hat{\mathbf{y}}^{(1)} \cdots \hat{\mathbf{y}}^{(G)}$라고 하겠습니다.
- 이때, $\hat{\mathbf{Y}} = \bar{\mathbf{X}}\hat{\mathbf{b}}$는 우리가 알고 싶어하는 M이라는 예측값 각각에 대한 G개의 부트스트랩 표본들의 결과를 보여줍니다. 따라서 $\hat{\mathbf{Y}}$는 $M \times G$의 행렬로 나타납니다.

8.0.3 비모수 부트스트랩을 통해 얻는 것 (3)

매번 부트스트랩 표본 추출을 해서 원하는 예측값을 저장하기보다, 그냥 $g \leq G$일 때, $\hat{\beta}^{(g)}$를 저장하면 됩니다. 간단하게 말하면, 부트스트랩으로 얻은 계수 추정치들만 가지고 있으면 된다는 얘기입니다. 그러면 그 계수 추정치, $\hat{\mathbf{b}}$를 우리가 원하는 부트스트랩 표본의 행렬($K \times G$)에 대입하여 계산하기만 하면 됩니다. 새롭게 다시 부트스트랩 과정을 반복할 필요 없이 이미 얻은 이 계수의 벡터를 이용해 자유롭게 예측값의 변화를 추적할 수 있습니다. 예를 들어, x_1과 x_2가 각각 성별과 소득이라고 할 때, 여성과 최저소득자로 고정해서 예측값을 구하고 싶다면? 다른 변수들은 부트스트랩 표본에서 그대로 놔두고 부트스트랩 표본에서 해당 두 열의 값만 각각 여성과 초저소득자에 해당하는 값으로 고정한 뒤, 이미 구한 계수의 행렬과 계산해주면 됩니다.

8.1 모수적 부트스트랩

자, 비모수 부트스트랩을 복습해보았습니다. 이제부터는 챕터 8의 주제인 모수적 부트스트랩에 대해 살펴보겠습니다. 다시 표집분포에 대한 이야기로 돌아가 보겠습니다. OLS에서 표집분포란 우리가 얻을 수 있는 모든 가능한 추정치 $\hat{\beta}$가 $(\beta, \hat{\sigma}^2(X'X)^{-1})$을 따르는 정규분포를 가진다는 것을 의미합니다. OLS가 BLUE를 산출할 수 있는 가정들을 만족시킨다면 말이죠. 그런데 문제는 우리가 β를 모른다는 것입니다. β는 모집단의 모수, 우리가 진정 알고자 하는 모집단에서의 관계를 보여주는 것이니까요.

King, Tomz, 그리고 Wittenberg (2000, 이하 King et al. (2000))는 왜 우리가 OLS를 통해 추정해낸 "최고의 선형관계를 보여주는 편향되지 않은 효율적인 추정값"(Best Linear Unbiased Estimator)인 $\hat{\beta}$를 사용할 것을 제안합니다. King et al. (2000)에 따르면, 시뮬레이션을 통해 구할 수 있는 모든 계수값은 OLS 계수 추정치를 평균으로 하고 표준오차에 따라 정규분포를 이룰 것이라고 기대할 수 있습니다: $\tilde{\beta}\ N(\hat{\beta}, \hat{\sigma}^2(X'X)^{-1})$. 만약 이같은 관계가 성립한다면 우리는 이 정규분포로부터 원하는 값을 자의적으로 선택해서 보여줄 수 있고, $\tilde{\beta}$에 대해 어떤 함수적 관계에 상관없이 계산을 할 수 있게 됩니다.

8.1.1 평균과 불확실성을 추정하기

실제 모델을 통해서 한 번 생각해보도록 하겠습니다. 이 모델은 임금수준과 교육수준, 그리고 그 두 변수의 상호작용을 통해 범죄율을 설명할 수 있다는 내용을 담고 있습니다. 정확히는 임금수준이 범죄율에 미치는 효과가

교육수준에 따라 조건적으로 나타날 것이라고 기대한다고 합시다.

$$\text{범죄율} = \beta_0 + \beta_1 \text{임금수준} + \beta_2 \text{교육수준} + \beta_3 (\text{임금수} \times \text{교육수준})$$

여기서 우리가 관심있는 것은 바로 관찰된 교육 수준의 값이 평균일 때, 임금 수준의 변화에 따른 기대값이라고 할 수 있습니다.

- 모든 교육수준 관측치에 대해 임금 수준을 고임금($\overline{\text{임금}}$)과 저임금($\underline{\text{임금}}$)의 저임금으로 벡터화해보겠습니다.
- $g = 1, \cdots .G$라고 할 때, 다음과 같이 단계대로 진행합니다.
    1. $\tilde{\beta}^{(g)}$ $N(\hat{\beta}, \hat{\sigma}^2 (X'X)^{-1})$의 분포를 따르는 $\tilde{\beta}^{(g)}$를 추출합니다.
    2. 교육수준 변수로부터 $PE^{(g)}$를 무작위로 추출합니다.
    3. $\mu\overline{\text{임금}}^{(g)} = \tilde{\beta}_0^{(g)} + \tilde{\beta}_1^{(g)} \overline{\text{임금}} + \tilde{\beta}_2^{(g)} \text{교육수준}^{(g)} + \tilde{\beta}_3^{(g)} (\overline{\text{임금}} \times \text{교육수준}^{(g)})$을 계산합니다.
    4. $\mu\underline{\text{임금}}^{(g)} = \tilde{\beta}_0^{(g)} + \tilde{\beta}_1^{(g)} \underline{\text{임금}} + \tilde{\beta}_2^{(g)} \text{교육수준}^{(g)} + \tilde{\beta}_3^{(g)} (\underline{\text{임금}} \times \text{교육수준}^{(g)})$을 계산합니다.
    5. $\mu\overline{\text{임금}}^{(g)}$과 $\mu\underline{\text{임금}}^{(g)}$ 값을 저장해줍니다.

자, 그러면 이제 두 개의 벡터, $\mu\overline{\text{임금}}^{(g)}$과 $\mu\underline{\text{임금}}^{(g)}$를 가지게 되었습니다. 이 두 벡터를 요약해서 보여주기만 하면 됩니다. 그러면 우리는 임금이 매우 낮을 때와 임금이 매우 높을 때의 범죄율의 차이를 분포를 통해 직관적으로 확인할 수 있습니다. 이 방법이 효율적인 이유는 위의 4번과 5번에서처럼 $\tilde{\beta}$의 G개의 추출 결과를 가지고 그냥 $\tilde{\mathbf{b}}$의 행렬과 결합한 뒤 계산만 해주면 되기 때문입니다. 아니면 그냥 바로 구한 $\tilde{\beta}$ 행렬을 요약해서 보여줘도 되구요.

8.1.2 King et al. (2000)

이 파트에서는 King et al. (2000)의 주장과 제안을 간단하게 정리 및 소개하도록 하겠습니다. 나아가 그들이 (1) 왜 이런 논문을 썼는지, (2) 예측값(predicted values)와 기대값(expected values)의 차이점이 무엇인지, 그리고 1차 차분(first difference)이란 무엇인지, (3) 베이지안 접근법에 대해서 King et al. (2000)이 무어라 말하고 있는지 등에 대해 살펴보겠습니다.

8.1.2.1 Title: Making the Most of Statistical Analyses: Improving Interpretation and Presentation

이 논문의 제목은 우리가 통계 분석을 통해 얻을 수 있는 결과들의 특성에 대해 좀 더 숙고할 필요가 있다는 King et al. (2000)의 제안을 드러낸다고 할 수 있습니다. 이들은 연구자들이 단지 복잡한 통계분석의 결과를 날것 그대로 보고하지만 말고, 그것의 의미를 제대로 전달할 수 있는 방식으로 보고할 책임이 있다고 주장하고 있습니다. King et al. (2000)에 따르면, 사회과학자들은 통계분석 결과로부터 "가용한 정보의 모든 장점을 거의 취하지 못하고" 있습니다(King et al. 2000: 347). 달리 말하면, 당시 King et al. (2000)이 정치학 분야의 통계방법을 이용한 경험연구들을 살펴보았을 때, 통계 결과를 보여주고 해석하는 방식이 실질적으로 그 결과를 이해하는 데 충분한 정보를 제공하지 못하고 있었다는 것을 의미합니다. 물론 지금은 이 논문이 작성된지 20년이 흘렀고, King et al. (2000) 이후로도 많은 발전과 변화가 있었습니다.

King et al. (2000)은 독자가 통계학적 훈련을 받지 않은 이라고 할지라도 연구자들은 그들이 읽을 수 있고, 이해할 수 있는 방식으로 통계 결과를 제시할 책임이 있다고 강조합니다. 또한 통계적 유의성이 실질적 유의성과 같은 것일고 할 수는 없기 때문에 우리는 연구 결과로 나타난 통계 결과의 실질적 의미가 무엇인지에 대해 통계적 바탕을 가지지 못한 다른 사회과학자 또는 논문을 읽을 일반 독자들과 공유하는 방법을 숙고해야 한다는 것입니다.

이들은 또 통계 방법이 가지는 두 가지 불확실성에 대해 인정해야만 한다고 언급하고 있습니다. 바로 추정결과의 불확실성(estimation uncertainty)과 본연적 불확실성(fundamental uncertainty)입니다. 이 두 불확실성으로 인해 통계적 결과는 가능한 한 많은 정보를 제공해야만 합니다. 단지 효과의 규모, 방향, 그리고 유의성만 보여줄 것이 아니라 우리가 추정해낸 이 효과가 얼마나 불확실한지 역시도 보여주어야 한다는 것입니다. 이를 위해서 King

et al. (2000: 350)은 우리가 관심을 가지고 있는 예측값, 기대값, 그리고 1차 차분값이라는 점 추정치에 대해 더 많은 정보를 제공할 수 있을 것으로 기대하는 시뮬레이션에 기반한 접근법을 제안하고 있습니다.

- King et al. (2000)이 말하는 예측값은 시뮬레이션 결과로 얻은 예측값을 의미합니다. 통계적 시뮬레이션은 표본 추출 횟수에 따라 수많은 시뮬레이션 표본을 산출할 수 있습니다. 이는 우리가 가지고 있는 통계 모델을 가지고 수없이 시뮬레이션을 돌려 시뮬레이션 결과값들을 가질 수 있다는 것을 의미합니다. 따라서 예측값은 이 시뮬레이션 결과로 가지게 된 벡터들로 계산된 결과입니다. 만약 우리가 예측값을 계산한다고 하면, 시뮬레이션으로 얻은 계수 값에 각 변수에서 하나의 관측치씩을 대입하여 하나의 결과값을 얻게 됩니다. 이것이 바로 예측값입니다. $y_1 = \beta_0 + \beta_1 \times (x_1 = 1) + \beta_2 \times (x_2 = 3) + \beta_3 \times \{(x_1 = 1) \times (x_2 = 3)\}$에서 결과로 나타난 이 y_1가 각각의 변수에 특정 값들을 대입했을 때, 우리가 얻을 수 있는 예측값입니다.
- 한편 기대값은 예측값의 변동성을 평균화한, 결과값의 평균값이라고 할 수 있습니다. 따라서 우리는 기대값이 예측값의 변동성의 평균을 취하기 때문에, 예측값이 기대값보다 더 큰 분산을 가질 것이라고 생각할 수 있습니다.
- 그리고 두 기대값 사이의 차이를 우리는 1차 차분값이라고 합니다. 연구자들은 예측변수들의 설정을 다르게 함으로써 서로 다른 두 개의 기대값을 얻는 알고리즘을 사용합니다(King et al. 2000: 351). 어려울 것은 없습니다. 이 전 파트에서 다른 변수인 교육수준은 똑같이 넣고 임금 수준만 고임금, 저임금으로 나누어서 결과의 변화를 살펴보았던 것이 바로 이런 설정입니다. 1차 차분값은 다른 변수들이 통제되었을 때, 우리가 관심을 가지고 있는 주요 변수의 변화가 결과에 미치는 효과를 보여준다고 할 수 있습니다. 따라서 우리는 이 알고리즘을 반복하면 1차 차분값의 분포를 얻을 수 있습니다. 이 분포를 이용해 우리는 1차 차분값에 대한 추정값과 표준오차를 구할 수 있게 됩니다. 모수적 부트스트랩에서 수많은 부트스트랩 표본들을 이용해 각각의 1차차분값들을 모아 분포를 보여주는 것과 같은 논리입니다.

King et al. (2000)은 베이지안 접근법에 대해서도 하나의 대안적 접근법으로 평가하고 있습니다. 이들은 베이지안 접근법이 중심극한정리나 정규성 가정과 같은 제약에서 상대적으로 자유롭다는 점에서 더 설명력 있는 결과를 제시할 수 있다고 제안합니다. 하지만 King et al. (2000)이 논문을 작성하던 시점에서는 베이지안 접근법을 기존의 통계적 기법들에 적용하기 어려웠기 때문에, 이들은 상대적으로 용이한 시뮬레이션 기반의 접근법을 사용할 것을 제안하고 있습니다.

2020년 7월 현재, 이 논문은 구글 스칼라에서 약 4,100여 명이 넘는 학자들에 의해 인용되고 있습니다. 물론 인용하는 이들이 모두 King et al. (2000)의 주장과 제언을 수용하는 이들ㅇ은 아닐 것입니다만, 적어도 이들의 주장이 한 번쯤 연구문제를 적실하게 풀어나가는 방법을 고민할 때, 되새겨볼만한 의미를 가진 논문이라는 것을 시사한다고 생각합니다.

8.1.2.2 Practice the King et al. (2000)!

머리 아픈 논문을 살펴보았으니, 이제 데이터를 이용해서 모델을 만들고 King et al. (2000)의 제안에 따라 그들의 논문에 있는 Figure 1과 같은 방식으로 결과를 재현해보도록 하겠습니다. 두 가지 그래프를 그려야 합니다. 이를 위해서 일단 추정해야 할 모델은 일반적인 형태로 다음과 같은 구조를 가지고 있다고 가정해보겠습니다.

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \beta_4 w + \beta_5 m + \beta_6 mw$$

하나의 그래프는 x와 z에 초점을 맞춘 것이어야 하고, 다른 하나는 m과 w의 관계를 보여주어야 하는 것이라고 하겠습니다. 먼저 `QoG` 데이터셋의 2015년도 자료를 통해서 변수들을 다음과 같이 모델링 해보겠습니다.

$$\begin{aligned} \text{경제규모} =& \beta_0 + \beta_1 \text{무역개방성} + \beta_2 \text{민주주의} + \beta_3 (\text{무역개방성} \times \text{민주주의}) \\ &+ \beta_4 \text{농지 비율} + \beta_5 \text{노령인구비율} + \beta_6 (\text{농지비율} \times \text{노령인구비율}) \end{aligned}$$

나중에야 `R`의 `Zelig` 패키지나 `STATA`의 `Clarify`를 이용할 수 있겠지만, 여기서는 기본적인 논리를 이해하며 코딩을 전개하고자 하기 때문에 좀 원시적인(?) 방식으로 코드를 짜보도록 하겠습니다.

```
pacman::p_load(ezpickr, mvtnorm, tidyverse)

QOG <- pick(file =
```

```
              "http://www.qogdata.pol.gu.se/data/qog_std_ts_jan20.dta")
QOG2 <- QOG %>%
  dplyr::filter(year == 2015) %>%
  dplyr::select(ccode, wdi_gdpcapcon2010, p_polity2, wdi_trade,
                    wdi_pop1564, wdi_agedr, wdi_araland) %>% drop_na()
```

데이터를 불러와 주었으니 이제는 모델을 만들어야겠죠? 단위의 문제가 있으니 종속변수인 경제규모의 측정지표, 1인당 GDP는 로그값을 취하도록 하겠습니다.

```
model1 <- lm(log2(wdi_gdpcapcon2010) ~ wdi_trade + p_polity2 +
               I(wdi_trade * p_polity2) + wdi_araland + wdi_agedr +
               I(wdi_araland * wdi_agedr), data=QOG2)
model1 %>% broom::tidy() %>%
  mutate_if(is.numeric, round, 3) %>%
  knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	16.851	0.715	23.576	0.000
wdi_trade	0.005	0.003	1.579	0.117
p_polity2	0.075	0.039	1.922	0.057
I(wdi_trade * p_polity2)	0.000	0.000	0.602	0.548
wdi_araland	-0.054	0.029	-1.845	0.067
wdi_agedr	-0.084	0.010	-8.705	0.000
I(wdi_araland * wdi_agedr)	0.001	0.000	1.131	0.260

일단 회귀분석 결과를 `summary`를 통해 살펴보면 상호작용들 모두 유의미하지 않는 것으로 나타나네요. 뭐 어쩔 수 없습니다. 그렇다고 하더라도 이 작업이 무의미하지는 않으니까요.

제가 관심이 있는 것은 변수들 간의 상호작용입니다. 먼저 x, z에 해당하는 변수, 무역개방성과 민주주의의 관계를 살펴보도록 하겠습니다.

```
set.seed(19891224)
beta_draws <- rmvnorm(n=1000, mean = coef(model1), sigma=vcov(model1))
head(beta_draws)
```

```
##      (Intercept)    wdi_trade  p_polity2 I(wdi_trade * p_polity2) wdi_araland
## [1,]    17.23158 -0.001321063 0.04673637             0.0005194318 -0.01716309
## [2,]    18.52625 -0.002157919 0.04129964             0.0008712935 -0.07170068
## [3,]    17.03413  0.001507908 0.01122512             0.0006596835 -0.05526781
## [4,]    18.07822  0.004928456 0.03685346             0.0006261671 -0.09405449
## [5,]    16.52870  0.007434208 0.10491406            -0.0002519455 -0.01721839
## [6,]    16.53057  0.005413327 0.13277980            -0.0001899500 -0.04771521
##        wdi_agedr I(wdi_araland * wdi_agedr)
## [1,] -0.08118915              -2.071477e-05
## [2,] -0.09922753               6.404959e-04
## [3,] -0.08193179               6.811284e-04
## [4,] -0.10168630               1.302814e-03
## [5,] -0.07824166              -1.239507e-04
## [6,] -0.08377743               4.760842e-04
```

`rmvnorm`은 다변량 정규분포(multivariate normal distribution)의 무작위 추출을 가능하게 하는 함수입니다. 위의 코드에 따르면 평균을 모델의 각 계수값으로 하고 각 계수들 간의 분산-공분산을 표준편차로 하는 분포로부터 각각의 계수값들의 1000회 추출한 시뮬레이션 분포를 `beta_draws`라는 객체에 저장하라는 내용입니다. 자세한 내용은 다루지 않겠습니다만, 간단하게 이야기하면 추정된 OLS 계수를 평균으로 하는 분포에서 시뮬레이션된 분포를 뽑되, 그 표준편차를 계수들 간의 공분산 즉, 계수들 간에 상관을 일종의 표준오차로 고려하여 추출하라는 것입니다.

이렇게 1000개에 해당하는 시뮬레이션된 계수값을 얻었으니, 이제는 우리가 관심을 가지고 있는 변수의 범위를 설정해주어야겠죠? 높은 수준의 무역개방성과 낮은 수준의 무역개방성을 보여주기 위하여 상위 15%에 해당하는 값과 하위 15%에 해당하는 값을 `quantile` 함수를 이용하여 벡터화 하였습니다. 이를 `trade`라는 별도의 객체에 저장하겠습니다. 또한 민주주의의 수준이 변화함에 따라 이 서로 다른 수준의 무역개방성이 경제규모에 미치는 효과를 탐색해야하기 때문에 민주주의 수준의 변화도 벡터화하여 포함하여 줍니다. `POLITY2`는 -10부터 10까지 1의 간격을 가진 변수기 때문에 그대로 반영해서 백터화하였습니다.

```
trade <- quantile(QOG2$wdi_trade, c(0.15, 0.85))
trade
```

```
##       15%       85%
##  45.28606 121.64338
```

```
democracy <- c(seq(-10, 10, by=1))
democracy
```

```
##  [1] -10  -9  -8  -7  -6  -5  -4  -3  -2  -1   0   1   2   3   4   5   6   7   8
## [20]   9  10
```

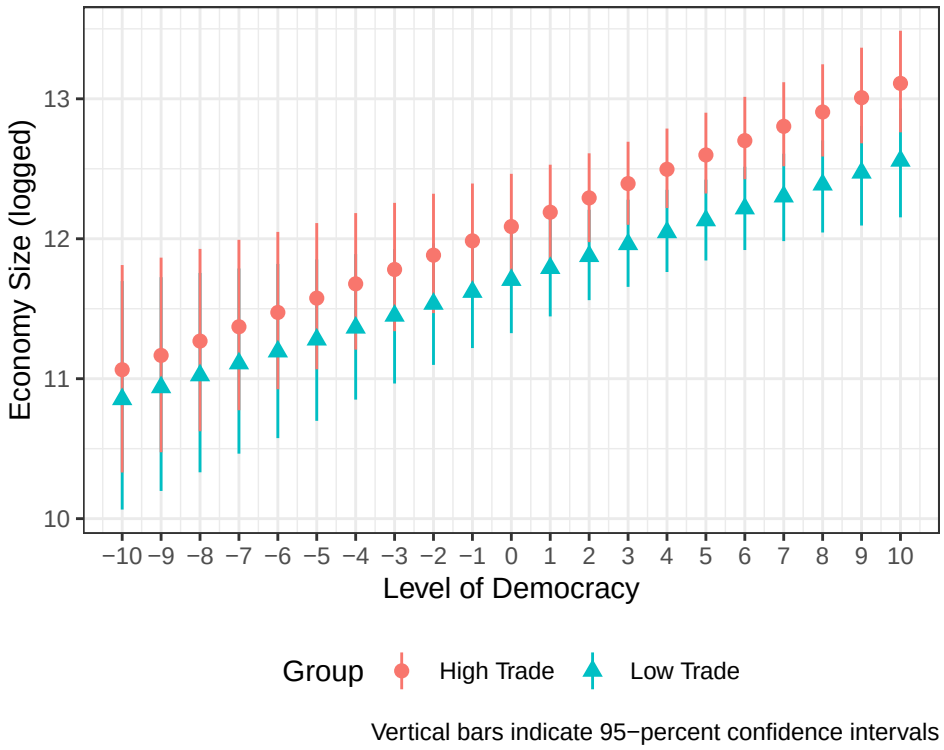
그리고 나서 이제는 $\bar{X}$를 무역개방성과 $\overline{\text{무역개방성}}$에 해당하도록 만들어줍니다. 무역개방성과 민주주의를 제외한 나머지 모든 변수들이 평균값을 가진다고 할 때, 무역개방성의 수준 차이가 민주주의 수준의 변화에 따라 어떻게 경제규모에 영향을 미치는지 살펴보겠습니다.

```
# 낮은 수준의 무역개방성
LowTrade <- cbind(1, trade[1],
                  democracy,
                  I(trade[1] * democracy),
                  mean(QOG2$wdi_araland), mean(QOG2$wdi_agedr),
                  I(mean(QOG2$wdi_araland) * mean(QOG2$wdi_agedr)))
# 높은 수준의 무역개방성
HighTrade <- cbind(1, trade[2],
                democracy,
                I(trade[2] * democracy),
                mean(QOG2$wdi_araland), mean(QOG2$wdi_agedr),
                I(mean(QOG2$wdi_araland) * mean(QOG2$wdi_agedr)))
```

나머지는 $\hat{Y} = \bar{X}\hat{b}$에 대한 항렬계산을 `R` 코드로 구현하는 것입니다.

```
LowTrade.ME <- t(LowTrade %*% t(beta_draws)) ## ME는 Marginal Effect입니다.
LT.mean <- apply(LowTrade.ME, 2, mean) ## 구해진 1,000개의 예측값의 평균
LT.se <- apply(LowTrade.ME,2,
               quantile, c(0.025, 0.975)) ## 구해진 1,000개의 예측값의 표준편차

HighTrade.ME <- t(HighTrade %*% t(beta_draws))
HT.mean <- apply(HighTrade.ME, 2, mean)
HT.se <- apply(HighTrade.ME, 2,
               quantile, c(0.025, 0.975)) ## 2.5/97.5 perentile

LT <- data.frame(Democracy=democracy,
                 Group = "Low Trade", ## 시뮬레이션이니 직접 하위 5%, 상위 5%의
                 # 관측치를 95% 신뢰구간을 위해 사용할 수 있습니다.
                 Mean=LT.mean, Lower=LT.se[1,], Upper= LT.se[2,])
HT <- data.frame(Democracy=democracy,
                 Group = "High Trade", ## 시뮬레이션이니 직접 하위 5%, 상위 5%의
                 # 관측치를 95% 신뢰구간을 위해 사용할 수 있습니다.
                 Mean=HT.mean, Lower=HT.se[1,], Upper= HT.se[2,])
Trade <- bind_rows(LT, HT)
```

```
Trade %>%
  ggplot(aes(x=Democracy, y=Mean, color=Group, shape=Group)) +
  geom_point() +
  geom_pointrange(aes(y = Mean, ymin = Lower, ymax = Upper)) +
  scale_x_continuous(breaks = democracy)+
  theme(axis.text.x  = element_text(vjust=0.5)) +
  labs(#title="Economy Size (logged) by the Level of Democracy",
       x="Level of Democracy", y="Economy Size (logged)",
       caption="Vertical bars indicate 95-percent confidence intervals") +
  theme_bw() + theme(legend.position = "bottom")
```



상호작용 변수가 유의미하지 않은 것을 쉽게 이해할 수 있습니다. 왜냐면 민주주의 수준이 변화하는 전 구간에 걸쳐서 높은 수준의 무역개방성과 낮은 수준의 무역개방성이 경제규모에 미치는 효과가 모두 중첩됩니다. 즉, 두 효과가 통계적으로 유의하게 차이난다고 볼 수 있는 경험적 근거가 충분치 않기 때문에 우리는 두 효과가 서로 같다(다르지 않다)라는 영가설을 기각할 수 없게 됩니다.

이번에는 농지 비율과 노령인구비율의 관계를 살펴볼 것인데요, 똑같습니다. 이번에는 농지비율의 변화에 따라 높은 수준과 낮은 수준의 노령인구비율이 경제규모에 미치는 효과의 변화를 살펴볼 것입니다. 다만, POLITY2와 다르게 농지 비율은 1 단위로 변화하지는 않습니다. 따라서 저는 summary 함수를 이용해 농지 비율 변수의 척도를 확인하고, 최소값부터 최대값까지를 포함할 수 있는 범주를 설정하고 5라는 단위로 인터벌을 가지게끔 하였습니다.

```
age <- quantile(QOG2$wdi_agedr, c(0.15, 0.85))
age
```

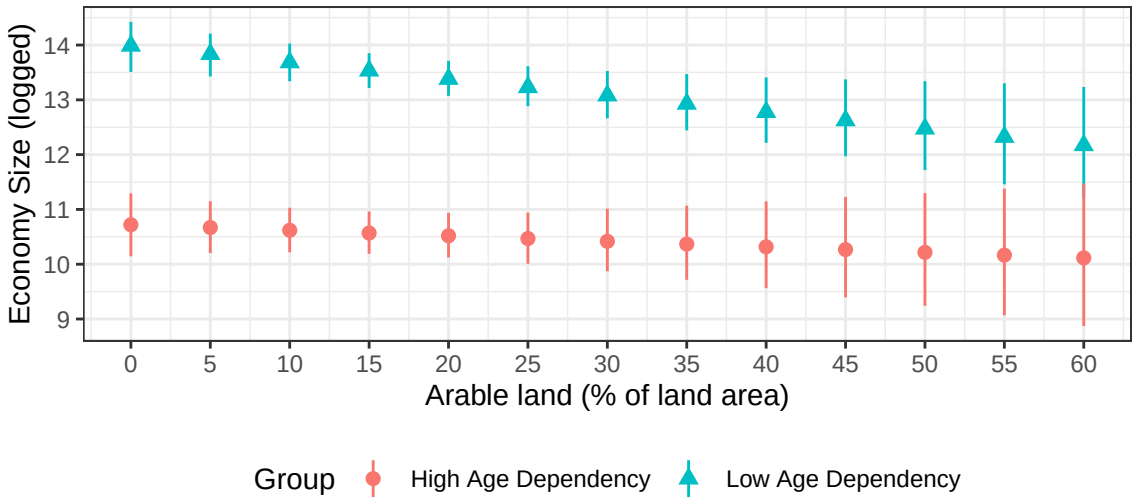
```
##      15%      85%
## 44.20432 82.99981
```

```
summary(QOG2$wdi_araland)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##  0.1685  5.1514 12.3059 16.1934 23.0050 59.4011
```

```r
agriland <- c(seq(0, 60, by = 5))


# 여기서 중요한 점은 예측변수의 행렬을 만들 때, 나중에 행렬곱셈을 해줄 시뮬레이
# 션된 계수값들의 순서는 OLS 분석에 투입된 변수 순서와 같다는 것입니다.
# 이를 고려해서 변수를 조작해주어야 합니다.

LowAge <- cbind(1, mean(QOG2$wdi_trade), mean(QOG2$p_polity2),
                I(mean(QOG2$wdi_trade) * mean(QOG2$p_polity2)),
                agriland, age[1],
                I(agriland * age[1]))

HighAge <- cbind(1, mean(QOG2$wdi_trade), mean(QOG2$p_polity2),
                I(mean(QOG2$wdi_trade) * mean(QOG2$p_polity2)),
                agriland, age[2],
                I(agriland * age[2]))

LowAge.ME <- t(LowAge %*% t(beta_draws))
LA.mean <- apply(LowAge.ME, 2, mean)
LA.se <- apply(LowAge.ME, 2,
                   quantile, c(0.025, 0.975))

HighAge.ME <- t(HighAge %*% t(beta_draws))
HA.mean <- apply(HighAge.ME, 2, mean)
HA.se <- apply(HighAge.ME, 2,
                   quantile, c(0.025, 0.975))

LA <- data.frame(Arable=agriland,
                 Group = "Low Age Dependency",
                 Mean=LA.mean, Lower=LA.se[1,], Upper= LA.se[2,])
HA <- data.frame(Arable=agriland,
                 Group = "High Age Dependency",
                 Mean=HA.mean, Lower=HA.se[1,], Upper= HA.se[2,])
Age <- bind_rows(LA, HA)

Age %>%
  ggplot(aes(x=Arable, y=Mean, color=Group, shape=Group)) +
  geom_point() +
  geom_pointrange(aes(y = Mean, ymin = Lower, ymax = Upper)) +
  scale_x_continuous(breaks = agriland) +
  theme(axis.text.x  = element_text(vjust=0.5)) +
  labs(#title="Economy Size (logged) by the Level of Arable land",
    x="Arable land (% of land area)", y="Economy Size (logged)",
    caption="Vertical bars indicate 95-percent confidence intervals") +
  theme_bw() +
  theme(legend.position = "bottom",
        legend.text=element_text(size=8.5))
```



어라? 여기서 재밌는 결과가 나옵니다. 틀림없이 농지 비율과 노령인구비율의 상호작용항은 통계적 유의성이 없는 것으로 나타났는데, 그래프는 농지 비율이 낮은 경우에는 낮은 수준의 노령인구비율이 경제규모에 미치는 효과가 높은 수준의 노령인구비율보다 더 높게 나타났습니다. 단, 이러한 차이는 농지 비율이 증가할수록 점차 수렴하여 종래에는 두 효과가 통계적으로 유의미한 차이를 보이지 않는 다는 것을 보여주었습니다.

이 결과가 보여주는 바는 명확합니다. 사실 OLS 결과표에서 상호작용항에 대한 계수값은 저 모든 효과를 뭉뚱그려서 평균 효과를 보여주는 것에 불과합니다. 아마도 노령인구비율이 고만고만한 차이를 보이는 중간 지점의 국가들에서는 저러한 효과가 눈에 보이지 않게 혼재되어 있을 수 있습니다. 하지만, 적어도 우리는 데이터를 통해서 1차 차분값의 결과가 저러한 정보를 내재하고 있다는 것을 보여줄 수 있고, 나아가 실질적인 함의를 도출하며 독자로 하여금 분석의 의의를 이해하는 데 도움을 줄 수 있습니다.

8.2 비모수 부트스트랩 vs. 모수적 부트스트랩

마지막으로 챕터 7과 여기 챕터 8에서 각각 다루었던 비모수 부트스트랩과 모수적 부트스트랩의 이론적 차이에 대해서 간단하게 얘기해보고자 합니다. 이 둘의 차이는 모집단의 분포를 가정하느냐, 혹은 가정하지 않느냐에 좌우된다고 할 수 있습니다. 비모수 부트스트랩이 모집단의 분포를 가정하지 않습니다. 비모수 부트스트랩을 사용해 우리는 하나의 표본(원 표본)으로부터 무수히 많은 수의 부트스트랩 표본을 추출할 수 있습니다. 이 부트스트랩 표본에 속하게 될 관측치들은 원 표본으로부터 모두 동일한 확률로 추출됩니다. 즉, 무작위로 추출됩니다. 비모수 부투스트랩은 부스트스랩 포본들을 통해 얻은 표본 통계치들의 분포를 표집분포로 활용합니다. 따라서 비모수 부트스트랩을 이용할 경우, 우리는 OLS의 고전적인 가정들에 굳이 집착할 필요가 없습니다. 반복된 표집을 통해 우리는 반복 추출된 부트스트랩 표본들의 통계치들의 분포를 직접 관측할 수 있기 때문에 오차분산의 정규성 등과 같은 가정을 할 필요가 없습니다. 부트스트랩하고, 표집을 반복하면 우리는 바로 $\hat{\beta}$의 분포를 얻을 수 있으니까요.

한편, 모수적 부트스트랩은 우리가 얻은 표본이 알려지지 않은 모수를 가진, 알려진 분포로부터 얻어졌다고 가정합니다. 모수적 부트스트랩에 따르면 우리는 원 표본이 특정한 분포를 따르는 모집단으로부터 추출되었다고 가정합니다. 따라서 우리는 특정한 부포를 가정하는 통계 기법과 방법을 이용하여 표본을 가지고 모집단의 모수를 추정할 수 있습니다. 만약 우리가 모집단이 어떠한 분포를 따른다는 가정 하에서 표본이 뽑혔다고 가정한다면, 우리가 표본으로부터 얻은 추정값 역시 어떠한 분포를 가정했을 때, 그리고 모수가 가질 것으로 여겨지는 조건들을 충족했을 때 신뢰할만한 것이라고 여길 수 있다는 것입니다. 예를 들어, OLS 추정치가 BLUE라고 주장할 수 있다고 하겠습니다. 그렇다면 우리는 이 계수추정치에 모수적 부트스트랩을 적용하기만 하면 됩니다. 중심극한정리에 따라, 부트스트랩 표본 추출 횟수가 증가할수록 OLS 계수값을 평균으로 취하는 표집분포는 정규 분포에 점차 근사할 것이기 때문입니다. 모수적 부트스트랩을 통해 우리는 OLS 추정결과에 대해 분석적 결과(analytic results)에 비해서 더 풍부한 정보를 가진 수치들을 얻을 수 있습니다.

만약 우리가 부트스트랩 표본 추출 횟수를 많이 할 수 있다고 한다면, 모수적 부트스트랩과 비모수 부트스트랩은

표본 규모가 작더라도 유사한 결과를 제공할 것입니다. 그러나 부트스트랩 표본 추출 횟수가 감소하거나 표본의 분포가 매우 한 쪽으로 치우쳐져 있다거나 극단적인 값들을 많이 포함하고 있다면, 이 두 부트스트랩 방법을 통해 얻은 결과는 꽤 다를 수 있습니다.

8.3 조금 더 나아가기: Hainmueller, Mummolo, and Xu (2019)

2019년에 `Political Analysis`에 게재된 Hainmueller, Mummolo, and Xu의 논문, "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice"는 King et al. (2000)의 근본적인 문제제기와 앞선 챕터들에서 살펴본 상호작용항을 단지 회귀분석 결과표의 β만 보고 섣부르게 판단해서는 안 된다는 문제 모두를 잘 다루고 있습니다. 이 논문은 정치학 분야에서 흔히 사용하는 상호작용항을 다루는 방식에 대한 문제를 제시하고 있습니다. 동시에 Brambor et al. (2006)에 대해서도 일종의 업데이트를 하고 있는 논문입니다. 한번쯤 꼭 읽어보시기를 권하고 여기서는 필요에 따라 간단하게 요약 및 정리하는 정도로 마무리하겠습니다.

8.3.1 Hainmueller et al. (2019)의 주장과 Brambor et al. (2006)에 대한 비판

Hainmueller et al. (2019)은 Brambor et al. (2006)가 곱셈을 통해 나타나는 상호작용항을 탐색하고 해석하는데 있어서 일종의 가이드라인을 제공하고는 있지만 몇 가지 중요한 이슈들을 간과하거나 언급조차 하고 있지 않다고 비판합니다. Hainmueller et al. (2019)에 따르면 그 문제는 크게 두 가지로 대별할 수 있습니다. 첫째, 선형 상호작용(linear interaction effect; LIE)에 대한 가정과 둘째, 충분한 정보량의 결여에 관한 것입니다.

Brambor et al. (2006)는 편미분을 취하는 방식을 통해서 상호작용항의 한계효과를 살펴보고 있습니다. 아래는 Hainmueller et al. (2019: 166)가 제시하고 있는 수식들로 고전적 선형회귀분석 모델에 곱셈 형태의 상호작용항이 포함된 모델을 보여줍니다.

$$Y = \mu + \eta X + \alpha D + \beta(D \cdot X) + Z\gamma + \epsilon \tag{8.1}$$

이 모델에서 Y는 종속변수이고, D는 우리가 관심을 가지고 있는 핵심적인 예측변수, 혹은 처치변수입니다. X는 일종의 매개변수이고, $(D \cdot X)$는 상호작용변수, Z는 일련의 통제변수들이라고 하겠습니다. μ, ϵ은 각각 상수와 오차항을 보여줍니다. 이때, 핵심적인 예측변수 D의 종속변수 Y에 대한 한계효과는 다음과 같이 나타낼 수 있습니다.

$$\text{ME}_D = \frac{\partial Y}{\partial D} = \alpha + \beta X \tag{8.2}$$

이 지점에서 Hainmueller et al. (2019)는 Brambor et al. (2006)이 간과한 점이 있다고 지적합니다. Brambor et al. (2006)의 논의에 따르면 우리는 단지 핵심 변수들의 한계효과가 일종의 선형 함수적 형태로 나타나리라고 가정해야 합니다. 그러나 Hainmueller et al. (2019)는 그 선형상호작용에 대한 가정은 결코 선험적인 것이 아니며 종종 유지되지도 않는다고 주장합니다. 따라서 Hainmueller et al. (2019)는 연구자들이 그들의 데이터를 한 번 더 살펴보고 한계효과를 진정 선형함수의 형태로 나타낼 수 있는지를 의심해보라고 주문합니다. Brambor et al. (2006)이 묵시적으로 LIE를 가정하고 많은 학자들이 그 가이드라인을 그저 따른다고 할지라도 Hainmueller et al. (2019)는 LIE는 결코 선험적으로 정당화될 수 없는 가정이며, 이를 확인하기 위해서는 연구자가 가진 데이터를 더 깊이 들여다보고 이해하는 것이 필요하다고 지적합니다.

이이서 Hainmueller et al. (2019)는 충분한 정보량의 문제에 대해서 지적합니다. 이 문제는 상호작용항의 효과를 살펴볼 수 있는 데이터의 범주 전반에 걸쳐서 실상 우리가 관측할 수 있는 가용성의 문제와 직결됩니다. 다르게 표현하자면, 만약 우리가 매우 치우친 형태의 분포를 가진 데이터를 가지고 있다면 그 데이터에서 매개변수의 한 값에서 우리는 한계효과가 명확하게 존재한다고 할만한 충분한 정보를 얻지 못할 수도 있기 때문입니다. 수학적, 행렬적 계산으로 도출해서 예측값의 변화를 보여줄 수는 있지만 과연 그것이 실제로는 존재하지 않는 데이터의 구간을 수리적 계산으로 그릴 뿐이라면? 한계효과에 대한 주장의 타당성에 의문을 제기할 수 있다는 것입니다. Hainmueller et al. (2019)는 논문 165쪽에서 두 가지 조건에 대해 명시하고 있습니다: "(1) 주어진 매개변수의

값에 대해서 X 값이 충분한 수의 관측치를 가지고 있어야 하며, (2) 그 매개변수의 값에서 핵심적인 예측변수, D의 변화가 존재해야 한다는 것"입니다. 특히 때로 매우 치우치거나 값이 일정하게 분포되어 있지 않은 자료를 사용하는 사회과학연구에서 이같은 노력을 주문하고 있습니다.

만약 주어진 매개변수의 특정한 값에 실질적으로 핵심적인 예측변수의 관측치들이 없거나, 거의 존재하지 않는다면 우리는 충분한 정보없이 한계효과를 그저 추정하는 것이 되고, 이를 Hainmueller et al. (2019: 165)는 "함수적 형태의 외삽 또는 내삽의 문제"라고 언급하고 있습니다. 정리하자면, Hainmueller et al. (2019)의 두 가지 핵심적인 주장은 첫째, Brambor et al. (2006)이 간과하거나 묵시적으로 가정하는 LIE의 문제를 고려할 것, (2) 상호작용 모델에 관한 이론적 이해와 우리가 사용하는 경험적 데이터 간의 간극을 좁힐 것으로 요약할 수 있습니다.

8.3.2 Hainmueller et al. (2019)의 대안

8.3.2.1 첫 번째 전략

Hainmueller et al. (2019)의 첫 번째 전략은 데이터가 LIE 가정을 충족시키는지를 진단해보자는 것입니다. 이들은 원 데이터의 산포도를 그려볼 것을 추천합니다. 그냥 추천하는게 아니라 한계효과의 LIE 가정과 매개변수의 각 데이터 포인트별 핵심 예측변수의 실제 관측치 분포를 살펴볼 수 있는 산포도를 그리는 방법을 제시합니다.

- 첫째, 핵심 예측변수가 이항변수일 경우, 핵심 예측변수에 따라 그래프를 두 개의 패널로 나눈 뒤 매개변수와 종속변수 간의 관계를 보여주는 산포도를 그려보라고 합니다.
- 둘째, 두 개의 선을 이 산포도에 더하는데, 하나는 상호작용 효과의 선형성을 가정하는 회귀선이고, 다른 하나는 일종의 가중치를 적용한 국소가중치 회귀선(locally weighted regression; LOESS)입니다.

예측변수의 값에 따라 나뉜 산포도의 각 패널에서 이 두 선을 비교함으로써, 우리는 LIE가 충족되는지 여부를 확인할 수 있습니다.

마지막으로 정보량의 문제에 있어서 Hainmueller et al. (2019)는 데이터에 충분한 관측치들이 존재하는지를 보여줄 수 있는 박스플롯을 제시하라고 제안합니다. 만약 핵심적인 예측변수가 연속형 변수라면 표본을 대강 비슷한 규모를 가진 세 개의 집단으로 분리하여 매개변수에 따라 낮은 수준의 X (first tercile), 중간 수준의 X (second tercile), 그리고 높은 수준의 X (third tercile)의 패널로 나타내라는 것입니다(Hainmueller et al. 2019: 170). 이산형 변수일 때와는 달리, 연속형 변수일 경우 우리는 Y에 대한 D의 관계가 이 세 집단의 매개변수 패널에서 회귀선과 LOESS곡선에 어떠한 차이를 보이는지를 살펴보라는 것입니다. 이를 통해 우리는 선의 기울기의 변화 또는 차이를 통해 Y에 대한 D의 관계가 서로 다른 수준의 X에 따라서 어떻게 달라지는지를 파악할 수 있게 됩니다.

8.3.2.2 두 번째 전략

두 번째 전략은 일종의 구간화를 통한 추정치(binning estimator)를 사용하라는 것입니다. 구간화는 매개변수의 특정 값의 구간을 포함하는 일련의 더미변수들을 말하는데, 얘를 들어 매개변수가 1부터 10까지라면 1부터 3까지 하나의 더미변수, 4부터 7까지 또 다른 하나의 더미변수, 그리고 나머지 값들을 마지막 더미변수 등에 담는 방식을 말합니다. 우리가 이항변수인 매개변수를 가지고 있다면, 한계효과를 계산하는 것은 쉽습니다. 0과 1을 각각 집어넣으면 되니까요. 하지만 연속형 매개변수를 가지고 있을 때는 특정한 매개변수의 값을 골라서 한계효과를 살펴보기가 쉽지 않고, 살펴본다 하더라도 정확한 변화를 잡아내기가 쉽지 않습니다. Hainmueller et al. (2019)는 매개변수를 크게 세 개의 구간으로 나누어서 더미변수의 형태를 취하게 하고 이를 통해 매개변수의 삼분위 범주값들을 보여주라고 제안합니다. 만약 핵심적인 예측변수가 각각의 구간화 변수와 상호작용한다면, 우리는 주어진 매개변수의 삼분위 구간에서 예측변수의 한계효과를 보여줄 수 있을 것입니다.

다음으로 각 구간화 변수를 대표할 수 있는 값을 특정하여 그 특정한 지점에서 예측변수의 효과를 평가하라고 제안합니다. 구간의 평균 혹은 중간값이 될 수 있겠죠? 마지막으로 Hainmueller et al. (2019)는 핵심 예측변수와 구간화 더미, 그리고 예측값에서 그 평가지점(아까 이야기한 구간화 변수를 대표할 수 있는 평균 혹은 중앙값과 같은 수치) 간 차이 간 상호작용 모델을 추정하라고 제안합니다. 일종의 3개 변수 상호작용인 것이죠: $D \times X - x_j \times G_j$. 여기서 j는 각 구간화 변수를 의미합니다. Hainmueller et al. (2019: 171-172)에 따르면 LIE 가정이 충족되고 충분한 양의 데이터가 상호작용을 지지한다면, 구간화 추정치는 주어진 $ME_X = \hat{\alpha} + \hat{\beta}X$라는 한계효과로

나타나는 표준 상호작용 모델의 한계효과 불편추정값으로 수렴할 것입니다. 나아가 구간화 추정치는 매개변수의 값에 기초하여 구축된 것이니만큼 구간화를 이용해 추정된 조건적 한계효과는 외삽과 내삽의 문제에 크게 왜곡될 일이 없습니다. 즉, 가능한 한 많은 데이터를 이용해서 추정을 하게 된다는 것입니다.

구간화 추정치 그 자체는 한계효과가 선형인지에 대한 여부를 알려주지는 못합니다만, 이것을 가지고 우리는 그래프 등을 그려봄으로써 한계효과가 각 구간에서 일관된 선형 관계로 증가하는지 아니면 특정 구간에서 널뛰는지를 살펴볼 수 있습니다. Hainmueller et al. (2019)는 Figure 2와 Figure 4(b)를 통해 구간화 추정치를 통해 LIE 가정과 정보량의 문제도 함께 살펴볼 수 있다고 주장하고 있습니다.