Chapter 5

Recap of Interactions and Hypotheses Tests

바로 전 챕터에서 마지막 부분에 변수들 간의 상호작용(interaction)에 대해서 살펴보았습니다. 이 챕터에서는 우리가 선형회귀모델에서 어떠한 방식으로 이와 같은 변수들 간 서로 연계된, 상호작용하는 관계를 포착하고자 하는지, 그리고 상호작용은 우리가 흔히 이야기하는 '가설검정'(hypotheses test)의 맥락에서 어떻게 이해할 수 있는지를 살펴보고자 합니다. 일단 간단히 하나의 다중선형회귀모델을 식으로 나타내 보겠습니다.

$$y = \alpha + \beta x + \gamma z + \nu$$

위의 식에서 y에 대한 x의 효과는 무엇일까요? 우리는 편미분(partial derivatives)을 이용해서 $\partial y/\partial x$, x의 한 단위 증가분에 따른 y의 변화로 이를 나타낼 수 있고, 위의 식에서 이는 β라고 할 수 있습니다. 여기까지는 이제까지 살펴본 선형회귀모델에서 계수(coefficients)가 의미하는 것과 같습니다.

그럼 조금 더 깊게 들어가보겠습니다. x의 한 단위 증가분에 따른 y의 변화, 즉 $\partial y/\partial x$이 고정적이지 않고 변화한다면 어떻게 될까요? 예를 들어,

$$\partial y/\partial x = \beta + \gamma z$$

라고 표현할 수 있다고 하겠습니다. 이 식이 의미하는 것은 무엇일까요? 여기서의 β는 무엇을 의미할까요? 지금 두 부분으로 나누어서 언급한 다중선형회귀모델과 그 선형회귀모델에서 한 변수의 종속변수에 대한 효과가 상수(constants)가 아니라 변수(variables)일 경우가 바로 선형회귀모델에서의 상호작용을 이해하는 핵심이라고 보시면 됩니다.

5.1 상호작용 (Interactions)

이번에는 좀 더 직관적으로 모델이 상호작용을 전제하고 있다는 것을 보여줄 수 있는 회귀식을 구성해보도록 하겠습니다.

$$y = \alpha + \beta x + \tau xz + \gamma + \nu$$

자, 맨 처음의 식과 달라진 부분이 보이시나요? 위의 식에서 β는 무엇을 의미할까요? 과연 우리가 x에 대한 y의 효과를 맨 처음의 식에서와 같이 β만을 가지고 충분히 이해할 수 있을까요?

바로 위의 식에서는 y에 대한 x의 효과를 β로만 이해할 수 없습니다. 정확히는 y에 대한 x의 효과는 근본적으로 제3의 변수, z와 얽혀 있습니다. 따라서 이때 우리는 y에 대한 x의 효과를 β 그 자체로만 가지고 해석할 수

없습니다. 만약 z가 0일 경우에만 x와 y의 관계에서 β가 단독으로 의미를 가지게 됩니다. z가 0이면 위의 식에서 τxz라는 항이 아예 사라지게 되니 고려할 필요가 없게 되는 것입니다.

5.2 상호작용의 이해: Brambor et al. (2006)

정치학 분야에서 상호작용에 관한 방법론을 다룰 때, 거의 필수적으로 읽고 넘어가는 논문을 간단하게 한 번 살펴보고자 합니다. Brambor, Golder, 그리고 Clark이 2006년도에 저술한 "Understanding Interaction Models: Improving Empirical Analyses"이라는 제목의 논문입니다. 우리말로는"상호작용 모델의 이해: 경험적 분석의 개선" 정도로 이해할 수 있겠네요.

이 논문의 요지는 간단합니다. 상호작용항(interaction term)을 다중선형회귀모델에 투입하여 변수들이 종속변수에 대해 상호작용하여 미치는 영향을 살펴보고자 하는 이른바 상호작용 모델의 연구가설은 '조건적인 가설'(conditional hypotheses)을 가지게 됩니다. 즉, 서로 상호작용하는 변수들은 각자에 대하여 '의존적인' 관계에 놓이게 된다는 것이죠.

Brambor et al. (2006)은 이와 같은 상호작용항의 존재는 사실 다중선형회귀모델을 구성하는 Gaus-Markov 가정 중 변수들 간의 독립성에 해당하는 부분을 침해하는 것이기 때문에 이에 대한 조심스러운 접근이 필요함에도 불구하고 상호작용 모델을 사용하는 연구들이 고민없이 두 변수를 단순히 곱한 채 모델에 투입하는 등의 행태를 보이고 있다고 지적합니다.[1]

Brambor et al. (2006)은 곱셈을 통해 만들어낸 상호작용항을 이용한 경험적 분석을 개선할 수 있는 몇 가지 방법들을 제안합니다. 이들은 정치학 분야에서 조건적 가설을 사용한 몇몇 연구들이 상호작용 모델을 구성하는 데 있어서 오류를 범해 그 연구결과로부터 도출된 추론이 잘못되었다는 사실을 지적합니다. 과학적 연구란 항상 '정답을 보여주는' 연구가 아니라 어디까지나 '반증가능한' 연구를 의미한다는 것을 생각해볼 때, 그리 놀라운 일은 아닙니다. 오히려 굉장히 정치학 분야에서 과학적 연구의 순기능이 작동한 사례라고 볼 수도 있을 것 같습니다.

Brambor et al. (2006)의 핵심적인 주장—즉, 상호작용 모델을 이용한 경험적 연구의 개선방안은 크게 네 가지 정도로 정리해볼 수 있습니다.

- 첫째, 조건적 가설을 수립하고 그것을 경험적으로 분석하는 데에는 상호작용 모델이 요구된다.
- 둘째, 상호작용 모델을 사용할 때, 반드시 구성요소가 되는 변수들의 항(constitutive terms, 이하 구성항)을 모델에 모두 투입해야 한다. 예를 들어, x, xz, z 라는 세 예측변수들로 y를 설명하고자 할 때, 우리의 관심사가 상호작용의 관계, 즉 xz의 효과라고 하더라도 그 상호작용항을 이루는 x와 z의 각 변수 또한 모델에 투입해주어야 한다는 얘기입니다.
- 셋째, 구성항들은 기존의 다중선형회귀모델의 가산적 관계(additive relationship)에서 예측변수들을 다루듯 해석하기가 어렵다는 것입니다. 여기서 가산적 관계란 '+', OR로 이루어진 관계로 각 변수들 간의 관계는 독립적이라는 가정이 성립된다는 것을 의미합니다. 그러나 위에서 잠깐 살펴보았던 것처럼 상호작용항의 경우에는 x의 y에 대한 효과가 z에 조건적이고, 반대로 z의 y에 대한 효과도 x에서 자유롭지 못하죠. 따라서 각 변수들이 서로 독립적일 것이라는 기대가 가능했던 기존의 다중선형회귀모델과는 달리 상호작용 모델에서 구성항들의 의미는 직관적으로 해석하기가 까다롭습니다.
- 마지막으로 넷째, 연구자들은 반드시 실질적으로 유의미한 한계효과(marginal effects)와 표준오차를 계산해야 한다고 제안합니다(Brambor et al. 2005, 64).

논문의 저자들은 그들의 주장을 이해하기 쉽게 $Y = b_0 + b_1X + b_2Z + b_3XZ + \epsilon$이라는 모델을 통해 풀어나갑니다.

[1] 상호작용항을 곱셈으로 표현하는 것은 논리적 근거를 가지고 있습니다. 예를 들어, 부울리안(Boolean)은 '+'를 OR,'×'를 AND로 표현하는데, 곱셈이란 두 조건이 동시에 존재하는 것을 의미합니다. 상호작용항 역시도 두 변수의 효과가 함께 종속변수에 영향을 미친다는 것을 보여주고자 하므로 곱셈의 형태로 나타냅니다.

5.2.1 왜 조건적 가설을 수립하고 검증하기 위해서는 상호작용 모델을 사용해야만 하는가?

이론적으로 가산적 관계를 상정한 선형 모델은 각각의 예측변수들이 서로 독립적으로 종속변수와 관계를 맺고 있을 것이라고 봅니다. 이는 선형 모델의 경우 각 예측변수에 대해 조건적 가설이 아닌 독립적 가설을 수립하고 있다고 이해할 수 있습니다. 논문에 보시면 흔히, "다른 조건들이 모두 일정할 때, x의 y에 대한 효과는 어떠할 것이다" 라는 형태의 진술을 가설로 사용하는 것이 그러합니다. 다른 조건들이 모두 일정하다는 얘기는 다중선형회귀모델에서 우리가 관심을 가지고 있는 x 이외의 다른 모든 변수들을 상수로 고정하였을 때, x가 y에 미치는 부분적 효과(partial effect)만을 보겠다는 것이고, 이것이 이 챕터 맨 처음에 제시하였던 식에서의 β라고 이해할 수 있습니다.

그러나 여기서는 정치학이라고도 할 수 있겠습니다만 사회과학 제분야의 경우 어떠한 맥락 혹은 조건의 효과를 고려하는 것이 모델링에 주요한 영향을 미치기 때문에 상호작용항을 포함한 모델을 생각해보아야 하는 경우가 생길 때가 많습니다. 예를 들어, 한국의 선거정치 속에 나타나는 계급투표에 대해 관심이 있다고 해보겠습니다. 이때, 기존 연구들의 주요 가설은 소득 수준에 따라 돈 많은 사람은 세금 많이 떼기 싫고 기득권 층일테니 보수정당을 지지할 것이라고 기대하고, 반대로 돈 없는 사람은 복지지출 등을 제고하기를 기대하며 진보정당에 투표할 가능성이 클 것이라고 해보겠습니다. 이때, 나의 연구가설은 이러한 계급적 투표가 실제 선거에서 정당 선택으로 이어지는 관계가 개별 유권자들의 정치적 세련도, 혹은 정치지식 수준에 따라 달라진다는 조건적 가설일 수 있습니다. 왜냐하면 돈이 없더라도 정치지식 수준이 높은 사람들의 경우 특정 정책에 대한 이해도가 더 높아 정당 선호가 달라질 수 있으니까요. 다양한 논의가 가능하겠지만 일단 이런 맥락에서 사회과학 분야에서 상호작용 모델은 상당히 빈번하게 논의됩니다.

마찬가지로 Brambor al. (2006)도 조건적 가설의 사례를 하나 제시하는데, 그들이 제시하는 가설은 조금 더 통계적이라고 해야할까요, 도식적입니다. 실제 연구문제를 바탕으로 한 가설은 아닙니다. 일단 조건적 가설이라고 보기 어려운 경우를 한 번 보겠습니다.

그들은 "Z라는 조건이 존재할 때, X의 한 단위 증가가 Y와 관계를 가지고, Z라는 조건이 부재할 경우에는 X와 Y 간의 관계 또한 성립되지 않을 것이다"라는 가설을 제시합니다. 즉, 여기서 Z라는 변수는 특정 조건의 존재 여부를 보여주므로 이항변수(binary variable)이라고 할 수 있겠죠? 만약 우리가 이항변수 Z를 성별이라고 한다면, 이 성별 변수의 계수값은 Y에 있어서 남성과 여성일 경우 각기 다른 절편값으로 해석할 수 있을 겁니다. 남성이 1, 여성이 0이라고 한다면

$$\text{여성} : Y = b_0 + b_1 X + b_2(Z = 0) + \epsilon = b_0 + b_1 X + \epsilon$$
$$\text{남성} : Y = b_0 + b_1 X + b_2(Z = 1) + \epsilon = b_0 + b_1 X + b_2 + \epsilon$$

이때 이 두 식의 차이는 b_2, 상수로 나타납니다. 따라서 이 경우에는 조건적 가설이라고 부르기가 어렵습니다.

5.2.2 왜 상호작용 모델에 모든 구성항을 다 포함해야할까?

그렇다면 왜 상호작용 모델에 모든 구성항을 다 포함해야만 할까요? 그리고 상호작용 모델에서의 구성항과 일반 가산적 관계의 선형 모델에서의 예측변수 간의 차이는 무엇일까요?

상호작용 모델($Y = b_0 + b_1 X + b_2 Z + b_3 XZ + \epsilon$)에서 구성항을 제외한다면, 우리는 $Y = b_0 + b_1 XZ + \epsilon$라는 모델을 생각해볼 수 있습니다. 구성항을 제외해도 말이 된다라는 얘기는 위의 $Y = b_0 + b_1 X + b_2 Z + b_3 XZ + \epsilon$이 $Y = b_0 + b_1 XZ + \epsilon$의 모델과 다르지 않다라는 주장을 해야한다는 것을 의미합니다. 이때, 만약 구성항을 제외한 모델이 타당하다면,

1. Z가 평균적으로 Y에 미치는 효과가 존재하지 않다거나
2. X가 0일 때, Z가 Y에 미치는 효과가 없다

라고 주장할 수 있어야만 합니다. 다시 말하면, 첫째로 Z가 평균적으로 Y에 미치는 효과가 없다는 얘기는 사실상 비체계적 요인으로서 Y에 평균적으로 체계적 변화를 가져오는 요인이 X 뿐이라고 주장하던가(이 경우 본질적으로 해당 모형의 함의는 $Y = b_0 + b_1 X + \epsilon$과 다를 바 없게 됨), 혹은 X가 0일 때, $b_1 XZ$ 항이 제거되면서 절편값인 $b_0 + \epsilon$만 남게 되어 두 번째 주장이 타당하다고 말할 수 있어야 합니다. 그러나 위의 두 주장은 가설로서 거의 정당화되기 어렵습니다(Brambor et al. 2006, 66). 왜냐하면 $Y = b_0 + b_1 X + b_2 Z + b_3 XZ + \epsilon$에서 b_2는 상호작용 모델에서 X가 0일때의 Z가 Y에 미치는 효과를 보여주기 때문에, 이는 Z가 평균적으로 Y에 미치는

효과가 존재않다는 것을 보여주어야 하는 것과 배치됩니다. 마찬가지로 Z의 Y에 대한 평균적인 효과가 0이라고 하더라도 b_2가 반드시 0이라는 보장도 없습니다. X가 0일 때, Z가 Y에 대해 가지는 효과가 전혀 없다($b_2 = 0$)고 미리 전제하는 것보다는, b_2가 0이 아닐 수도 있을 가능성을 먼저 생각해보는 것이 더 많은 경우의 수를 제공합니다 (Brambor et al. 2005, 66). 다시 말해, 선험적으로 구성항을 제외하는 분석은 b_2가 0이 아니라면 우리가 관심을 가지고 있는 모수의 추정치를 왜곡하여 잘못된 추론을 도출하게 할 수 있습니다(Brambor et al. 2006, 68).

위에서 언급한 바와 같이, X에 대한 계수, b_1와 Z에 대한 계수 b_2는 일반적인 가산적 관계의 모델($Y = b_0 + b_1X + b_2Z + \epsilon$)의 X와 Z의 계수들과는 다릅니다. 가산 모델에서 b_2는 Z의 Y에 대한 평균적인 효과를 보여줍니다. 그러나 상호작용 모델에서 b_2는 X에 따라 조건적으로 변화하는 Z의 Y에 대한 효과를 보여줍니다.

5.2.3 왜 실질적으로 유의미한 한계효과와 표준오차를 보여주어야 하는가?

상호작용 모델에 있어서 각 예측변수들이 종속변수에 미치는 효과를 제대로 보여주기 위해서 Brambor et al. (2006)은 실질적으로 유의미한 한계효과와 표준오차를 함께 제시해야 한다고 주장하고 있습니다. 그 이유는 상호작용항의 효과가 일정하지 않기 때문(not constant)입니다. 상호작용 모델($Y = b_0 + b_1X + b_2Z + b_3XZ + \epsilon$)에서 X의 한 단위 증가와 관계된 Y의 변화분을 살펴보기 위해 편미분을 할 경우, 우리는 아래와 같은 식을 얻게 됩니다.

$$\frac{\partial Y}{\partial X} = b_1 + b_3Z$$

위의 식은 상호작용 모델에서 X의 Y에 대한 효과가 Z 변수의 값에 따라서 조건적으로 변화한다는 것을 의미합니다 (Brambor et al. 2006, 73). 따라서 상호작용항의 계수는 직접적으로 해석될 수는 없습니다. 그 값이 Z에 따라 변화하니까요. 우리는 계수의 부호나 통계적 유의성에 대해서는 이야기할 수 있지만 실질적으로 그 효과의 크기 등에 대해서는 계수만 보고는 대답할 수 없게 됩니다.

때문에 Z에 의해 조건적으로 변화하는 Y에 대한 X의 효과를 살펴보기 위해서는 X의 Y에 대한 한계효과를 계산할 필요가 있습니다. 한계효과는 Z의 조건 하에서 X의 한 단위 변화가 평균적으로 얼마만큼의 Y의 변화와 관계되는지를 보여줍니다. 또한 한계효과의 표준오차는 그렇게 계산된 한계효과가 얼마나 확실한지를 보여줍니다. 한계효과의 표준오차가 가지는 함의는 일반적으로 선형회귀모델의 계수값과 표준오차 간의 관계와 비슷하다고 이해하셔도 무방합니다. 그 한계효과가 통계적으로 유의미한 효과인지를 보여주는 것이니까요.

정리하자면 Brambor et al. (2006)의 함의는 다음과 같습니다.

- 상호작용 모델이라고 하더라도 상호작용항뿐만 아니라 모든 구성항을 포함시켜 분석해야 한다.
- "Z를 상수로 고정한다(=통제한다)"는 것은 Z가 0이라는 것과 같은 의미가 아니다.
- 모델 내에 존재하는 두 구성항의 곱으로 상호작용항을 만들었기 때문에 다중공선성(multicollinearity)가 발생할 수 있다.[2]
- 상호작용항의 계수는 $\frac{\partial Y}{\partial X} = b_1 + b_3Z$로 나타낼 수 있고, 이때 상호작용의 계수가 편미분을 하더라도 Z라는 새로운 변수에 의해 조건적으로 변화할 수 있다는 것을 알 수 있다. 따라서 상호작용항의 계수를 직접적으로 일반선형회귀모델의 계수처럼 해석하기는 어렵다.

그렇다면 한 번 실제 데이터를 통해 분석해보도록 하겠습니다.

5.3 상호작용항의 이해: 경험적 분석

4개의 연속형 변수(x, z, y, w)를 가지고 다음과 같은 형태의 모델을 추정하고자 한다고 합시다: $y = \gamma + \eta x + \alpha z + \mu xz + \beta w$. x와 z의 상호작용항이 포함되었으니 위의 식에서 x와 z의 효과를 단순히 η와 α를 가지고 해석할 수는 없을 것입니다. 따라서 여기서는 x와 z의 변화에 따라서 각각 z와 x가 y에 대해 미치는 효과가 어떻게 조건적으로 변화하는지를 살펴보고자 합니다.

[2] 다중공선성이라는 것이 변수들 간의 공변 양상(covariances)에 따라 나타나는 것이니만큼, 두 변수의 곱을 통해 만들어낸 상호작용항이 모델 전반의 다중공선성을 높일 것이라 예상하는 것은 어렵지 않습니다. 다만, 과연 다중공선성이 반드시 나쁘냐하는 것에 대해서는 고민해볼 필요가 있습니다. 앞선 챕터 4에서 이에 관한 문제를 다룬 부분이 있으니 참고해보시기 바랍니다.

- 분석을 위해서 변수 x와 z가 높은 수준의 값을 지니는 경우를 $\bar{x}$와 $\bar{z}$라고 하겠습니다.
- 반대로 x와 z가 낮은 수준의 값을 지닐 때를 $\underline{x}$, $\underline{z}$라고 표현하도록 하겠습니다.

그리고 위와 같은 분석적 프레임 하에서 다음과 같은 경우를 실제 모델과 데이터를 통해 살펴보도록 하겠습니다.

$$E(y|x = \underline{x}, z = \underline{z}) - E(y|x = \bar{x}, z = \underline{z})$$
$$E(y|x = \underline{x}, z = \bar{z}) - E(y|x = \bar{x}, z = \bar{z})$$

x와 z에 관한 위의 두 표현이 과연 모델과 데이터와 관련해서 우리에게 어떤 실질적 함의를 제공해줄까요?

먼저 QOG 데이터셋에서 2014년도의 국가코드, 국가명, 연도, 해외직접투자 유입, 노동가능 인구, 무역 개방성, 1인당 GDP에 해당하는 변수들을 따로 선별하여 서브셋을 만들고, 결측치를 제외하였습니다. 그리고 1인당 GDP를 해외직접투자 유입, 무역 개방성, 그리고 해외직접투자 유입과 무역 개방성의 상호작용항과 노동가능 인구로 설명할 수 있다는 모델을 만들었습니다.

```
library(ezpickr)
library(broom)
library(tidyverse)
QOG <- pick(file = "http://www.qogdata.pol.gu.se/data/qog_bas_ts_jan20.dta")
QOG.s <-QOG %>%
  select(ccode, cname, year,
         wdi_fdiin, wdi_pop1564, wdi_trade, wdi_gdpcapcon2010) %>%
         # 이 네 변수들은 모두 연속형 변수들입니다.
  dplyr::filter(year==2014) %>% drop_na()

model.ch5 <- lm(wdi_gdpcapcon2010 ~ wdi_trade + wdi_fdiin +
                  wdi_trade*wdi_fdiin + wdi_pop1564, data=QOG.s)
```

원래는 모델에서 얻은 추정치(estimates)를 별도의 객체(objects)로 저장하여 작업하는 것을 선호하지는 않습니다. 하지만 여기서는 직관적으로 위에 써 놓은 두 모델과 연관지어 이해하기 위해서 각 변수와 계수를 별도의 객체로 저장한 뒤에 위의 두 식과 동일한 R 코드를 이용해 분석해보도록 하겠습니다.

```
# 먼저 모델에서 얻은 각 계수값을 위의 식에 상응하는 객체로 저장하여 줍니다.
b0 <- model.ch5$coefficients[1]
b1 <- model.ch5$coefficients[2]
b2 <- model.ch5$coefficients[3]
b3 <- model.ch5$coefficients[4]
b4 <- model.ch5$coefficients[5]

# 그리고 각 변수들도 별도의 객체로 저장해보도록 하겠습니다.
y <- QOG.s$wdi_gdpcapcon2010; x <- QOG.s$wdi_trade;
z <- QOG.s$wdi_fdiin; w <- QOG.s$wdi_pop1564

# 이제 첫 번째 수식에 대응하는 계산을 수행합니다.
y1 <- b0 + (b1 * min(x)) + (b2 * min(z)) +
  (b3 * min(x) * min(z)) + (b4 * w)
y2 <- b0 + (b1 * max(x)) + (b2 * min(z)) +
  (b3 * max(x) * min(z)) + (b4 * w)
A1 <- y1 - y2
unique(A1)
```

```
## [1] 1756329 1756329
```

```
# 두 번째 수식에 대응하는 계산을 수행합니다.
y3 <- b0 + (b1 * min(x)) + (b2 * max(z)) +
  (b3 * min(x) * max(z)) + (b4 * w)
```

```
y4 <- b0 + (b1 * max(x)) + (b2 * max(z)) +
  (b3 * max(x) * max(z)) + (b4 * w)
A2 <- y3 - y4
unique(A2)
```

```
## [1] -12068740 -12068740
```

위에서 만든 모델은 x와 z, 즉 해외직접투자 유입과 무역 개방성 간의 상호작용항을 포함하고 있습니다. 그리고 이 두 변수는 '서로에 대해 조건적'입니다. 따라서 두 변수가 모두 그 값이 변화한다고 할 때, 우리는 각 변수의 높은 값과 낮은 값을 최대값, 최소값으로 생각하여 다음과 같은 네 가지 시나리오를 생각해볼 수 있습니다.

$$\begin{aligned}
\hat{y_1} &= \gamma + \eta\underline{x} + \alpha\underline{z} + \mu\underline{x}\underline{z} + \beta w \\
\hat{y_2} &= \gamma + \eta\bar{x} + \alpha\underline{z} + \mu\bar{x}\underline{z} + \beta w \\
\hat{y_3} &= \gamma + \eta\underline{x} + \alpha\bar{z} + \mu\underline{x}\bar{z} + \beta w \\
\hat{y_4} &= \gamma + \eta\bar{x} + \alpha\bar{z} + \mu\bar{x}\bar{z} + \beta w
\end{aligned}$$

위의 네 시나리오에 따라서 우리는 $\hat{y_1} - \hat{y_2}$과 $\hat{y_3} - \hat{y_4}$를 계산해볼 수 있습니다.

$$\begin{aligned}
\hat{y_1} - \hat{y_2} &= (\hat{\gamma} + \hat{\eta}\underline{x} + \hat{\alpha}\underline{z} + \hat{\mu}\underline{x}\underline{z} + \hat{\beta}w) - (\hat{\gamma} + \hat{\eta}\bar{x} + \hat{\alpha}\underline{z} + \hat{\mu}\bar{x}\underline{z} + \hat{\beta}w) \\
&= \hat{\eta}(\underline{x} - \bar{x}) + \hat{\mu}(\underline{x}\underline{z} - \bar{x}\underline{z}) = \hat{\eta}(\underline{x} - \bar{x}) + \hat{\mu}\underline{z}(\underline{x} - \bar{x}) \\
\hat{y_3} - \hat{y_4} &= (\hat{\gamma} + \hat{\eta}\underline{x} + \hat{\alpha}\bar{z} + \hat{\mu}\underline{x}\bar{z} + \hat{\beta}w) - (\hat{\gamma} + \hat{\eta}\bar{x} + \hat{\alpha}\bar{z} + \hat{\mu}\bar{x}\bar{z} + \hat{\beta}w) \\
&= \hat{\eta}(\underline{x} - \bar{x}) + \hat{\mu}(\underline{x}\bar{z} - \bar{x}\bar{z}) = \hat{\eta}(\underline{x} - \bar{x}) + \hat{\mu}\bar{z}(\underline{x} - \bar{x})
\end{aligned}$$

복잡해 보이지만 위의 식은 서로 상쇄되는 항들을 정리하면 다음과 같은 두 개의 식으로 다시 쓸 수 있습니다.

$$\begin{aligned}
\frac{\Delta y}{\Delta x}\text{Given } \underline{z} &= \hat{\eta} + \hat{\mu}\underline{z} \\
\frac{\Delta y}{\Delta x}\text{Given } \bar{z} &= \hat{\eta} + \hat{\mu}\bar{z}
\end{aligned}$$

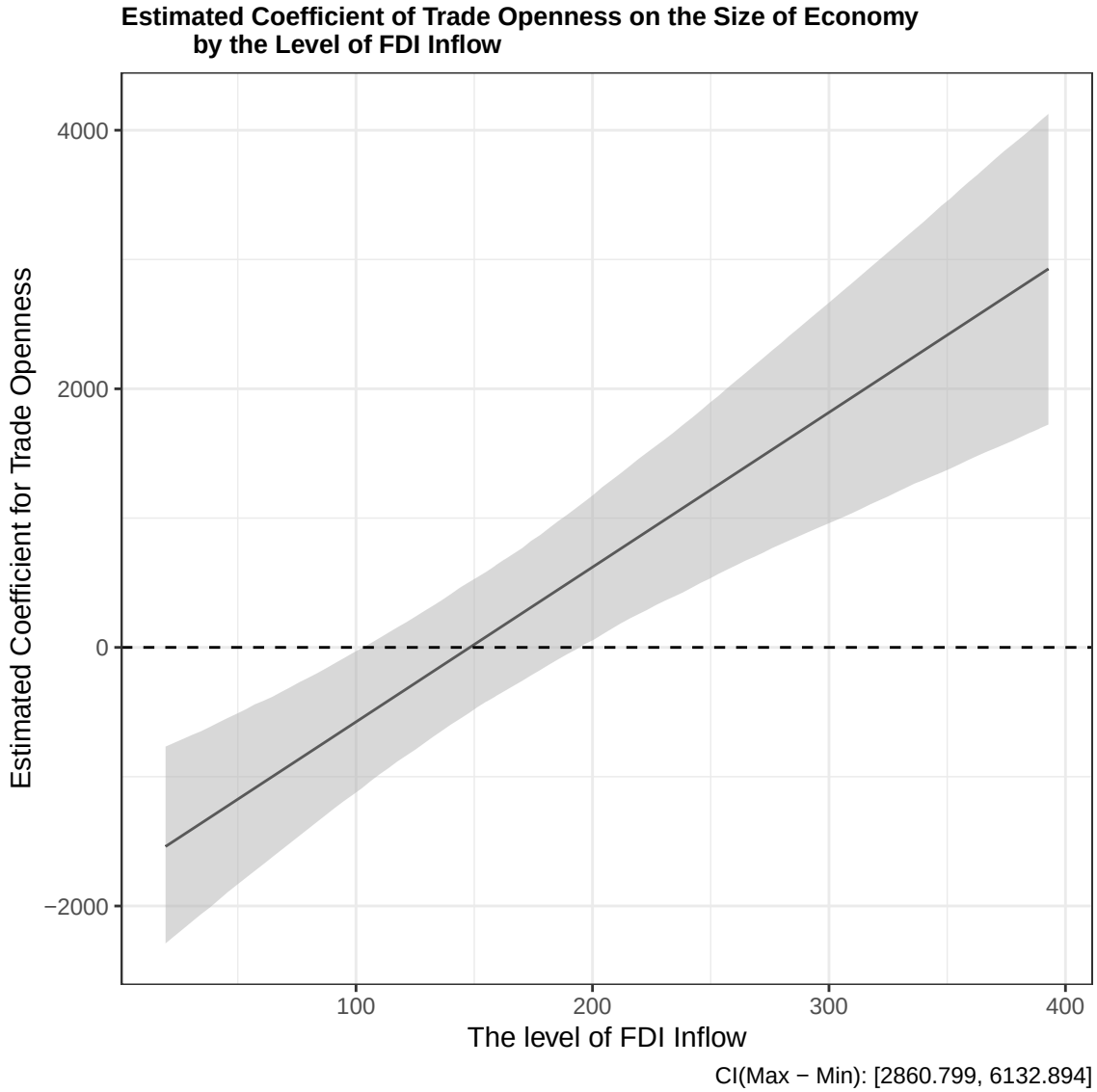
첫 번째 식, $\hat{y_1} - \hat{y_2}$을 통해 우리는 z가 낮은 값으로 고정(통제)되어 있을 때의 y에 대한 x의 효과, 기울기를 구할 수 있습니다. 반대로 두 번째 식, $\hat{y_3} - \hat{y_4}$를 이용해서 z가 높은 값으로 일정할 때, y에 대한 x의 효과를 추정할 수 있습니다. 실질적으로 데이터를 통해 구성한 위의 모델은 "한 국가의 경제수준 $= \gamma + \eta$무역 개방성 $+$ α해외직접투자 유입 $+ \mu$무역 개방성 $\times$ 해외직접투자 유입 $+ \beta$노동 가능인구로 다시 쓸 수 있습니다. 따라서 위의 모델을 통해 저는 무역 개방성이 경제수준에 미치는 효과는 해외직접투자 유입 수준에 따라 조건적일 것이라고 기대한 모델을 구축한 것입니다.

하지만 해외직접투자 유입이 연속형 변수이기 때문에, 해외직접투자 유입이라는 조건 하에서 무역 개방성이 경제수준에 미치는 한계효과를 포착하기란 쉽지 않습니다. 해외직접투자 유입이라는 변수의 값이 고정되어 있는 것이 아니라 변화하니까요. 그러므로 위에서는 해외직접투자의 유입에 관해 임의의 두 값, 최대값과 최소값을 설정함으로써 무역 개방성이 한 단위 증가할 때, 해외직접투자 유입의 최대값, 또는 최소값 하에서 경제수준에 미치는 한계효과를 계산하고자 한 것입니다. 만약 해외직접투자 유입의 최대-최소값으로 고정된 무역 개방성의 한계효과의 차이가 확실하게 보인다면, 이 모델을 통해 x와 z, 무역 개방성과 해외직접투자 유입 간의 상호작용 효과가 존재한다고 말할 수 있습니다.[3]

[3] 여기서 한 가지 생각해보고 넘어가야할 것이 있습니다. 물론 상호작용 모델은 앞선 챕터들에서 보았던 가산적 관계의 선형회귀모델과는 다르게 두 변수 간의 상호작용을 모델에 반영하고 있습니다. 하지만 편미분을 통해 살펴본 결과는 상호작용 모델 역시 그 내부에 선형함수를 포함하고 있다는 것입니다. $\frac{\partial y}{\partial x} = \beta_1 + \beta_3 z$라는 결과는 결국 상호작용 관계도 z에 따라 선형으로 나열된다는 것입니다. 그렇다면 만약 상호작용 효과가 비선형 관계라면 어떻게 될까요? 이에 관한 부분은 뒤에 다루도록 하겠습니다. 상호작용항을 모델에 포함할 경우, 각 변수들의 독립성을 가정한 가산적 모델보다 상대적으로 다양한 분석을 수행할 수 있게 되는 것은 맞지만 어디까지나 상호작용항도 일련의 가정에 기초하고 있기에 만능은 아니라는 점을 알아두어야 합니다.

R에서는 이와 같은 상호작용의 효과를 직관적으로 이해할 수 있도록 그래프를 통해 보여주는 여러 패키지들을 제공합니다. 직접 계산해서 ggplot으로 그려주셔도 좋습니다. 저는 후자를 선호합니다만, 여기서는 간단하게 패키지를 이용하여 위의 분석을 그래프로 재현해보도록 하겠습니다.

```
library(margins)
library(interplot)
interplot(m = model.ch5,
          var1 = "wdi_fdiin",
          var2 = "wdi_trade") +
  xlab("The level of FDI Inflow") +
  ylab("Estimated Coefficient for Trade Openness") +
  theme_bw() +
  ggtitle("Estimated Coefficient of Trade Openness on the Size of Economy
          by the Level of FDI Inflow") +
  theme(plot.title = element_text(face="bold", size = 10)) +
  geom_hline(yintercept = 0, linetype = "dashed")
```



이 그래프는 해외직접투자 유입 수준이 증가할수록 무역 개방성이 경제수준(경제규모)에 미치는 효과가 증가한다는

것을 보여주고 있습니다. 위에서 계산한 것은 해외직접투자 유입이 최소값이었던, 위의 그래프에서 제일 좌측의 값과 해외직접투자 유입이 최대값이없던 최우측의 값이라고 이해할 수 있습니다.

5.4 소결: 상호작용과 가설검정

이 다음 챕터에서 가설검정과 유의수준, 통계적 유의성에 관한 이야기를 다루고자 합니다. 그 전에 간단하게 상호작용과 가설검정 간의 관계를 짚고 넘어가고자 합니다. 기초 통계학을 공부하셨다면, 선형회귀모델에서 계수값을 통해 우리가 가설을 검정하는 방식에 대해서 이미 알고 계실 것입니다. 우리는 모집단에서의 모수들의 관계를 추론하기 위해 표본의 통계치들 간의 관계를 가지고 그 관계가 확률적으로 얼마나 '오류가 날지' 즉, 유의미하지 않은 관계일지를 통해 가설을 기각 혹은 기각하지 못합니다.

간단히 말하자면 표본은 본질적으로 모집단에서 추출해 모집단을 대표적으로 보여줄 것이라 기대되지만 표본추출의 방법 등에 내재된 한계로 인해 표본은 모집단과 동일할 수는 없습니다.

하나의 모집단에서 이론적으로 우리는 수없이 많은 표본들을 뽑아낼 수 있고, 이 표본들은 각각 평균과 같은 통계치를 가집니다. 따라서 우리는 표본들 통계치가 가지는 분포, 포집분포(sampling distribution) 등을 확인하게 되는 것이죠. 표본들이 문제없이 잘 뽑혔다면, 그리고 관측치의 수가 충분하다면 우리는 모집단의 기대값(expected value)이 표집분포의 평균에 수렴할 것이라고 기대하게 됩니다(중심극한정리).

하지만 어디까지나 표본은 모집단과 동일하지 않기 때문에 확률적으로 표본을 통해서 관측한 통계치들 간의 관계가 모집단에서 모수의 관계를 보여주지 못할 수도 있습니다. 보여주지 못할 확률이 우리가 설정한 어떠한 기준보다 클 경우 우리는 표본을 통해 분석한 결과가 통계적으로 유의미하지 않다(정확히는 유의미하다고 말하기 어렵다)고 결론을 내리게 됩니다. 이 점에서 선형회귀분석에서 x의 계수값 b_1은 모집단에서의 모수가 가지는 β_1를 보여줄 것이라는 기대를 가지고 있는 것입니다.

이때, 우리의 기대를 연구가설이라고 하면 이에 대한 영가설(null hypothesis)는 이러한 관계가 '존재하지 않을 것', 즉 $\beta_1 = 0$이라고 할 수 있습니다. 선형회귀분석에서 계수의 효과와 가설검정의 관계는 전체 표본 중에서 얼마나 많은 표본들이 관측된 결과가 0, 즉 "효과 없음"이라고 나타나느냐에 달려있다고 볼 수 있습니다.

그렇다면 상호작용 효과는 어떨까요? $y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon$이라는 모형이 있다고 할 때, 과연 상호작용 효과에 관심이 있기 때문에 β_3에만 관심을 가지고 $\beta 3 = 0$에 대한 영가설을 기각하면 될까요?

이것이 Brambor et al. (2006)이 유의미한 한계효과와 표준오차를 계산해야 한다고 했던 이유이기도 합니다. 왜냐하면 단지 $\beta_3 = 0$에 대한 기각 여부는 상호작용 효과를 이해하는 데 실질적으로 도움이 되지 않기 때문입니다. 상호작용 효과는 편미분을 했을 때, $\frac{\partial y}{\partial x} = b_1 + b_3 z$로 나타낼 수 있고 따라서 우리는 $\beta_3 = 0$이냐가 아니라 $b_1 + b_3 z = 0$인지를 살펴보아야 하기 때문입니다. z가 변수이므로 계속 변화하기 때문에 이 변화하는 $b_1 + b_3 z$, 한계효과와 그 표준오차를 계산해 그것이 얼마나 효과 없음, 0에 수렴하는지 혹은 수렴하지 않는지를 살펴보아야 한다는 것이 Brambor et al. (2006)의 핵심 주장입니다.