Chapter 4

Basics for Advanced Linear Regression Models: Part II

단순선형회귀모델은 단순해보이지만 결코 단순하지 않습니다. 앞에서 우리는 +가 각 변수들 간의 독립적인 관계를 보여준다는 것을 확인하였습니다. 이번에는 범죄에 대한 회귀모델을 구축하되, 임금 수준 이외의 또 다른 변수—경찰의 노력(Police effort)를 포함해봅시다. 선형회귀모델이되 둘 이상의 설명변수를 포함하였으니 다중선형회귀모델(multiple linear regression model)이라고 할 수 있을 것입니다. 통계적 모델로 나타내자면 범죄 $= \beta_0 + \beta_1$임금 수준 $+ \beta_2$경찰의 노력 $+ u$가 될 것입니다.

우리는 일단 다중선형회귀모델의 각 계수, β_i들을 어떻게 구할 수 있는지를 살펴볼 것입니다. 그리고 설명변수로서의 제곱항(square term)이 포함된 다중선형회귀모델에서 β_i가 가지는 의미를 분석하고자 합니다. 다중선형회귀모델 통계적 특성은 다음과 같습니다.

- MLR.1: 모수들은 선형관계를 이루고 있다.
- MLR.2: 표본은 무작위로 추출된 것이다(random sampling)
- MLR.3: 완전한 다중공선성은 존재하지 않아야 한다(No perfect collinearity)
- MLR.4: 오차항과 설명변수들 간의 상관관계가 존재하지 않아야 한다($E(u|x_i, \cdots, x_k) = 0$).

우리는 MLR.1부터 MLR.4까지를 가정하고, $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$를 가정합니다. 따라서 다중선형회귀모델에서 $\hat{\beta}_1$을 구하기 위해서는 다음과 같은 공식이 성립합니다.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} \cdot y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

이때, r_{i1}은 잔차로 $x_{i1} = \hat{\alpha}_0 + \hat{\alpha}_1 x_{i2} + r_{i1}$에서 도출된 것입니다. 간단히 말하자면, 단순선형회귀모델에서와는 달리 다중선형회귀모델에서는 $\hat{\beta}_1$을 구할 때, x_1과 y 간의 관계뿐 아니라 x_1과 x_2 간의 관계, 그리고 x_2와 y의 관계도 고려해주어야 한다는 것입니다. 왜? x_1이 설명하고 있는 부분의 일부는 사실 x_1과 x_2가 같이 설명하는 "교집합", "공분산"(covariance)일 수 있기 때문입니다.

공분산에 대한 내용은 뒤에서 더 자세하게 다루기로 하고, 우선 MLR.1부터 MLR.4까지의 가정들이 모두 충족되었을 때, 우리는 수없이 많은 표본들로부터 얻어내는 $\hat{\beta}_k$의 기대값이 모집단에서의 β_k와 같을 것이라고 생각할 수 있습니다: $E(\hat{\beta}_k) = \beta_k$. 이를 최소자승법으로 구한 선형회귀모델의 계수값의 비편향성(unbiasedness)라고 합니다.

4.1 오차(Errors)

이번에는 오차에 대해서 배워보겠습니다. 앞서 우리는 모집단 수준에서의 오차의 기대값은 0으로 수렴한다($E(u) = 0$)는 것을 가정하였습니다. 그렇다면 오차의 분산($Var(u)$)은 어떨까요? 이 오차의 분산에 관한 내용이 바로 다중선형회귀모델의 다섯번째 통계적 특성이자 가정, MLR.5입니다.

- MLR.5: 오차의 분산은 동질적이다(Homosckedasticity).
  - $Var(u|x_i, \cdots, x_k) = \sigma^2$
  - $E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
  - $Var(y|x_1, x_2) = Var(y) = Var(u) = \sigma^2 = E(u^2)$
  - 오차의 표준편차는 $\sqrt{\sigma^2} = \sigma$가 됩니다.
- 위에서부터 순서대로 저 가정의 내용을 풀어가보면, 먼저 우리는 주어진 설명변수들 하에서 오차항의 분산을 모집단 수준에서 σ^2라고 합니다. σ가 표준편차고 분산은 표준편차의 제곱이라는 수리적 정리를 여기서 딱히 증명할 필요는 없을 거 같으니 이렇게 넘어가겠습니다.
- 그리고 x_1, x_2라는 다중선형회귀모델의 두 설명변수가 있다고 할 때, 그 설명변수들이 주어졌을 때의 종속변수를 예측할 수 있는 기대값은 오차를 제외한 두 설명변수의 다중선형회귀모델의 PRF로 결정됩니다.
- 그리고 이러한 설명변수들이 주어졌을 때의 종속변수의 분산은 MLR.5에 따라 오차항의 분산과 같고, 오차항의 분산은 σ^2입니다. 이는 u^2의 기대값과도 같습니다.
- 표준편차는 분산의 제곱근이므로 결과적으로 오차의 표준편차는 σ가 됩니다.

4.1.1 $\hat{\sigma}^2$ 추정하기

표본에서의 오차—즉, 잔차의 분산($\hat{\sigma}^2$)은 표본의 크기의 영향을 받습니다. 정확히는

$$\hat{\sigma}^2 = \frac{1}{n-K-1} \sum \hat{u}_i^2$$

이며, 이는 곧 잔차의 제곱합을 n-k-1, 전체 관측치의 개수에서 변수의 개수-1을 하는 것과 같습니다($SSR/(n-K-1)$).

MLR.1부터 MLR.5까지의 가정이 모두 충족된다면 모집단 오차의 분산, σ^2에 대한 불편추정량인 $\hat{\sigma}^2$을 얻을 수 있습니다.

4.2 계수(Coefficients)

최소자승법(ordinary least square)는 예측값과 실제 관측치 간의 차이인 잔차의 제곱합을 최소로 하는 $\hat{\beta}_k$를 구하는 방법을 말합니다. 따라서 최소자승법에 따라 구한 β_i가 편향되어 있지 않다는 것은 모집단에서 추출한 표본들로부터 구한 각 $\hat{\beta}_k$의 기대값이 모집단의 β_k와 같다는 것($E(\hat{\beta}_k) = \beta_k$)을 의미합니다.

그러나 $\hat{\beta}_k$는 하나의 추정치에 불과하므로 필연적으로 불확실성을 내포하고 있습니다. 따라서 표본의 통계치로 모집단의 모수를 추론할 때에는 단지 $\hat{\beta}_k$를 제시하는 것만으로는 부족합니. 예를 들어, 우리는 "모집단에서의 계수, β_k가 0보다 큰가?"라는 질문에 대답하기 위해서는 다음과 같은 것들을 필요로 합니다.

- β_k에 대한 (최선의) 추정치인 $\hat{\beta}_k$
- 우리가 그 최선의 추정치에 대해 얼마만큼 "신뢰할 수 있느냐"에 관한 $\hat{\beta}_k$의 분산($Var(\hat{\beta}_k)$). 이 분산은 이론적으로 모집단에서 수많은 표본들을 뽑고 그 표본들에 대한 SRF에서 도출된 $\hat{\beta}_k$들이 얼마나 서로 다른지를 보여줍니다.

- 또, $\hat{\beta}_k$에 대해 신뢰하기 위해서 `t-statistic`, p-값 등과 같은 통계치들을 확인하고는 합니다.

4.3 분산(Variance)

분산에 관한 내용들이 계속해서 나오는데, 과연 분산이 크다는 것과 작다는 것은 무엇을 의미하는 것일까요? 다중선형회귀모델에서 분산이 중요하게 논의되는 맥락을 이해하기 위해서는 먼저 표집분포(sampling distribution)에 대해 짚고 넘어갈 필요가 있습니다.

4.3.1 표집분포(Sampling distribution)

$\hat{\beta}_k$을 확률변수라고 생각해봅시다. 주어진 하나의 표본에 대해서 $\hat{\beta}_k$는 고정된 값입니다. 그 하나의 표본에서 $\hat{\beta}_k$는 정해져 있기 때문입니다. 그러나 사실 우리는 모집단으로부터 또 다른 표본을 확보할 수 있습니다. MLR.2의 무작위 표집(혹은 확률표집) 가정에 따라서 우리는 하나의 모집단에서 뽑아낸 수없이 많은 표본들로부터 $\hat{\beta}_k$들을 일종의 확률변수로서 분포로 보여줄 수 있게 됩니다. 우리는 이 표본들로부터의 얻어낸 $\hat{\beta}_k$들의 분포를 표집분포(sampling distribution)라고 합니다. 그리고 우리는 그 표집분포가 모집단에 대한 β_k의 기대값을 포함하고 있을 것이라고 기대하고, $\hat{\beta}_k$의 분산—$Var(\hat{\beta}_k)$이 그 분포가 모집단의 β_k을 포함할 "불확실성"을 보여주는 것입니다.

4.3.2 표집분포의 표준오차(Standard errors)

여기서 우리는 수리통계적으로 $\hat{\beta}_k$에 대한 표준편차의 추정치에 대한 정리를 생각해볼 수 있습니다. 즉, $\hat{\beta}_k$의 표준오차에 대한 추정치는 표본에 따라 조건적이며, 동시에 MLR.1과 MLR.5 가정에 기초하고 있습니다.

$$se(\hat{\beta}_k) = \frac{\hat{\sigma}}{[SST_k \cdot (1 - R_k^2)]^{1/2}}$$

이때, SST_k는 x_k의 변동량을 의미하며, R_k^2는 다른 모든 설명변수들로 x_k를 회귀분석한 R^2 결과라고 할 수 있습니다.

표집분포의 표준오차는 또 다른 방식으로도 보여줄 수 있는데, 수리적으로만 유도해보도록 하겠습니다.

$$\begin{aligned} sd(\hat{\beta}_k) &= \sqrt{Var(\hat{\beta}_k)} \\ &= \frac{\sigma}{[SST_k \cdot (1 - R_k^2)]^{1/2}} \\ se(\hat{\beta}_k) &= \frac{\hat{\sigma}}{[SST_k \cdot (1 - R_k^2)]^{1/2}} \\ &= \frac{\hat{\sigma}}{\sqrt{n} \cdot sd(x_k) \cdot \sqrt{1 - R_k^2}} \end{aligned}$$

즉, 표준오차가 작을수록 우리가 표본을 통해 얻어낸 표본의 $\hat{\beta}_k$들이 만들어내는 분포가 모집단의 β_k을 포함하고 있을 가능성이 높다고 할 수 있습니다.

4.4 Best Linear Unbiased Estimator

그렇다면 이제는 이른바 BLUE, 불편추정량에 대해 알아볼 차례입니다. 대체 이 불편추정량이라는 것이 무엇을 의미하는 것일까요? 바로 모든 표본들을 통틀어 '평균적으로' 그 추정량이 최선의(효율적이고 편향되지 않은) 추

정량이라는 것을 의미합니다. 그리고 표준오차가 작다는 것은 그만큼 우리의 추정량이 "효율적"(efficient)이라는 것을 말합니다. MLR.1부터 MLR.5까지의 가정이 충족되는 하에서 OLS 추정량은 불편추정량입니다.

4.5 모델의 특정(specification)

4.5.1 부적절한(irrelevant) 변수의 포함

만약 우리가 추정해야할 모델이 $y = \beta_0 + \beta_1 x + u$라고 합시다. 그런데 모델을 $y = \beta_0 + \beta_1 x + \beta_2 z + e$로 수립해 추정하였다고 합시다. 이 경우에 부적절한 z 변수를 모델에 포함하여 모델을 잘못 특정하였다고 말합니다(misspecified).

4.5.1.1 잘못 특정된 모델: $E(\hat{\beta}_1)$과 $Var(\hat{\beta}_1)$에 미치는 영향

좀 더 자세히 이 잘못 특정된 모델이 $E(\hat{\beta}_1)$에 미치는 효과를 살펴보도록 하겠습니다. 우리는 MLR.1로부터 MLR.4까지의 가정 하에서 $E(\hat{\beta}_k) = \beta_k$가 된다는 사실을 알고 있습니다. 그렇다면 β_2, 즉 잘못 집어넣은 이 변수의 계수값이 0이라면 그것이 MLR.1부터 MLR.4까지의 가정에 영향을 미칠까요?

- 부적절한 변수를 모델에 포함하게 되면, 이 모델에서 우리는 $\tilde{\beta}_1$를 얻게 됩니다. $\hat{\beta}_1$이 제대로 된 모델에서 얻을 수 있는 OLS 추정치라고 합시다. 그러면 $E(\tilde{\beta}_1 | x_k) = \beta_1$라고 표현할 수 있습니다.
- 이 경우에 분산은 다음과 같은 관계를 가지게 됩니다.

$$\begin{aligned} Var(\hat{\beta}_1 | x_k) &= \frac{\sigma^2}{\sum_{i=1}^{N}(x_{i1} - \bar{x_1})^2} \\ &\leq \frac{\sigma^2}{(1 - R_1^2)\sum_{i=1}^{N}(X_{i1} - \bar{x_1})^2} = Var(\tilde{\beta}_1 | x_k). \end{aligned}$$

- $se(\hat{\beta}_1) = \frac{\sigma^2}{[SST_1 \cdot (1-R_1^2)]^{1/2}}$라는 것을 다시 한 번 떠올려봅시다.
  - 부적절한 변수를 모형에 포함시키더라도 SST_1는 변하지 않습니다.
  - 이때, R_1^2은 x_1을 종속변수로 하는 x_2의 회귀분석으로부터 얻어낸 R^2입니다.
    - 정확히는 $x_1 = \alpha_0 + \alpha_1 z + e$
  - 만약 z가 x를 잘 설명한다면, R_1^2는 높아질 것이고, $(1 - R_1^2)$는 작아질 것입니다. 따라서 분모가 작아지므로 결과적으로 $se(\hat{\beta}_1)$은 커지게 됩니다.
- 따라서 표본에서 x_1과 x_2가 서로 상관되어 있다면, x_2를 포함하는 것은 β_1에 대한 추정량의 분산을 증가시킬 수 있습니다.

결론적으로 부적절한 변수가 모델에 포함될 경우, 모수 추정에 편향성(bias)는 나타나지 않지만 모수에 대한 추정치의 표준오차가 증가하는 결과가 나타납니다.

4.5.2 적절한(relevant) 변수의 제외

한편, 우리가 원래 추정해야할 모델이 $y = \beta_0 + \beta_1 x + \beta_2 z + u$라고 해보겠습니다. 그런데 모델을 잘못 특정해서 $y = \beta_0 + \beta_1 x + e$로 추정했다고 할 때, $e = \beta_2 z + u$라고 할 수 있습니다. 이 경우에는 무엇이 문제일까요? MLR.1부터 MLR.5까지의 가정들이 위배되었다고 할 수 있을까요?

일단 x와 z의 상관관계가 없다, 즉 $Cov(x,z)=0$이라고 생각해보겠습니다. 이 경우에는 아무 문제가 없습니다. 왜냐하면 우리가 추정해야 하는 정확한 모델의 β_1와 잘못 특정한 모델의 β_1가 동일하기 때문입니다. 다만 잘못 특정한 모델의 오차항의 크기가 클 뿐입니다. 왜냐하면 오차항으로부터 y를 설명할 수 있는 z라는 요인을 추출해 내지 못했기 때문입니다.

문제는 바로 $Cor(x,z)\neq 0$일 경우입니다. Wooldridge(2016: 89-90)에서도 살펴볼 수 있듯이, 우리는 네 가지 누락변수로 인한 편의(Omitted variable bias, OVB)를 생각해볼 수 있습니다.

- 만약 $Cov(x,z)>0, \beta_2>0$이면, $\hat{\beta}_1>\beta_1$이므로 이 경우는 Positive OVB라고 합니다.
- 만약 $Cov(x,z)>0, \beta_2<0$이면, $\hat{\beta}_1<\beta_1$이므로 이 경우는 Negative OVB라고 합니다.
- 만약 $Cov(x,z)<0, \beta_2>0$이면, $\hat{\beta}_1>\beta_1$이므로 이 경우는 Negative OVB라고 합니다.
- 만약 $Cov(x,z)<0, \beta_2<0$이면, $\hat{\beta}_1<\beta_1$이므로 이 경우는 Positive OVB라고 합니다.

만약 적절한 변수를 제외하고 만든 모델로 OLS 추정을 하게될 경우에 우리는

$$\tilde{\beta_1}=\hat{\beta}_1+\frac{\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)e}{\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)^2}$$

를 얻게 됩니다. 즉, 원래 추정하고자 했던 $\hat{\beta}_1$에 비해 뭔가가 더 붙는 거슬 확인할 수 있습니다. 그렇다면 이 우변의 기대값을 구해보도록 하겠습니다.

- 일단, 위에서도 살펴보았던 것처럼 $e=\beta_2 z+u$입니다.
- 이걸 방금 전의 OLS 추정치, 우변에다가 대압해보겠습니다. 정말 끔찍한 수식이 나오지만 원래 추정하고자 했던 $\hat{\beta}_1$에 뭔가 점점 잡다한게 더해지고 있다는 것에서부터 이게 문제가 있다는 게 짐작이 가시겠죠?

$$\tilde{\beta_1}=\hat{\beta}_1+\frac{\hat{\beta}_2\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)x_{i2}+\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)e}{\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)^2}$$

이 수식으로 주어진 설명변수들에 대한 기대값을 구해보면,

$$\begin{aligned}E(\tilde{\beta_1}|x_k)&=\hat{\beta}_1\\&\quad+\frac{\hat{\beta}_2\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)x_{i2}+\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)E(e|x_k)}{\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)^2}\\&=\hat{\beta}_1+\hat{\beta}_2\frac{\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)x_{i2}}{\sum_{i=1}^{N}(x_{i1}-\bar{x}_1)^2}\\&=\hat{\beta}_1+\hat{\beta}_2\widehat{Cov(x_1,x_2)}/\widehat{Var(x_1)}\end{aligned}$$

라는 결과를 도출하게 됩니다. 따라서, 적절한 변수를 모델에서 제외할 때 생기는 편향, 누락변수의 편향(OVB)의 크기는

$$\text{Bias}(\tilde{\beta_1})=E(\tilde{\beta_1})|x_k-\beta_1=\beta_2\frac{\widehat{Cov(x_1,x_2)}}{\widehat{Var(x_1)}}$$

가 됩니다. 그렇다면, 위에서 살펴본 것 같이 두 경우를 생각해볼 수 있겠죠?

- $\hat{\beta}_2=0$
- $\widehat{Cov(x_1,x_2)}=0$

이 경우들에서는 편향성이 0이 됩니다. 따라서 일반적으로 종속변수에 영향을 미치는 변수를 누락시키는 문제는 누락변수가 포함된 변수들과 독립적이지 않은 한에야 포함된 변수들의 OLS 추정값들의 편향시키는 결과를 가져옵니다. 그리고 그 편향의 방향성과 크기는 위에서 살펴본 네 가지 경우의 수에 따라서 나타날 수 있으며, $\hat{\beta}_2$와 $\widehat{Cov(x_1, x_2)}$의 부호와 크기에 따라 결정됩니다.

4.6 다양한 변수들 (Various variables)

4.6.1 질적 변수(Qualitative variables)

`Lv.1.Statistics`에서 변수의 종류에 대해서 다루어본 적이 있습니다. 간략하게 말하면 일정한 간격을 가진 연속적인 수로 이루어진 연속형 변수(continuous variables)와 각 부류 간에 서로 배타적인 명목형 변수(혹은 분류형 변수, nominal or categorical variables), 그리고 그 사이에 서로 다른 부류임에도 불구하고 순위를 매길 수 있는 순위형 변수(ordinal variables)라고 구분할 수 있습니다.[1]

일단 변수가 연속형—양적 변수라고 할 때, 우리는 설명변수 x가 $\mathbb{R}$, 실수에 속한다고 할 수 있습니다. 즉 $x \in \mathbb{R}$이라고 한다면, $y = \alpha + \beta x + u$에서 β가 의미하는 것은 무엇일까요?

한편, 변수가 질적 변수—예를 들어, 명목형이라고 하겠습니다. 따라서 이때 새로운 변수 w는 어떤 분류목에 속한다는 의미에서 $w \in \{A, B, C\}$라고 하겠습니다. 이때의 $z = \gamma + \delta w + \nu$에서 δ가 의미하는 것은 무엇일까요?

위의 두 모델은 같은 형태를 가지지만 설명변수의 유형이 다릅니다. 양적 변수는 $x \in \mathbb{R}$이었고, 질적 변수는 $w \in \{A, B, C\}$로 나타낼 수 있었습니다. 바로 위에서는 모델의 특정(specification)에 대하여 살펴보았다면, 이번에는 설명변수의 유형에 따라 우리가 얻는 β_i의 의미가 어떻게 달라지는지를 살펴보고자 하는 것입니다.

질적 변수로 우리가 모델에 종종 포함하는 것은 이른바 더미변수(dummy variables, dummies)라고 불리는 것들입니다. 더미변수는 분류형 변수를 각 카테고리별로 쪼개어 각각 변수로 만든 결과로서, 그 카테고리에 속할 경우 1, 속하지 않을 경우 0으로 보여줍니다. 더미변수라고 부르는 이유는 이 변수가 해당 관측치가 특정한 카테고리에 속해있다 혹은 속해있지 않다는 것만을 말해주나는 점에서 그 정보량이 연속형이나 다른 유형의 변수들에 비해 상대적으로 적다는 것 때문입니다. w 변수를 예로 들자면, 우리는 w를 각각 $\{A, B, C\}$라는 카테고리별로 더미변수 3개로 쪼갤 수 있습니다.

$$w = \begin{cases} A & \text{iff } wA = 1 \\ B & \text{iff } wB = 1 \\ C & \text{iff } wC = 1 \end{cases}$$

그렇다면 이 더미변수들을 질적 변수로 구성했던 $z = \gamma + \delta w + v$에 투입하여 이 회귀모델을 실제 분석가능한 모델로 재구성해보도록 하겠습니다.

$$\begin{aligned} z &= \gamma_1 I(w = A) + \gamma_2 I(w = B) + \gamma_3 I(w = C) + \nu \\ z &= \gamma_1 wA + \gamma_2 wB + \gamma_3 wC + \nu \end{aligned}$$

자, 이렇게 분류형 변수를 더미변수들로 나누어 모델에 집어넣어 보면, 이론적으로는 문제가 없어보입니다. 왜냐하면 각각의 더미변수들을 모두 합쳐놓은 것이 원래의 분류형 변수일테니까요. 그럼 이대로 분석이 가능할까요? 아닙니다. 왜냐하면 위의 회귀분석모델은 **완전다중공선성(Perfect Mulitcollinearity)**로 인해 분석이 이루어지지 않게 됩니다. 세 더미변수의 총합, 즉 절편값이 1이 되기 때문입니다($wA + wB + wC = 1$). 따라서 모든 조건이 일정하다고 할 때, $z = \gamma + \delta_1 wA + \delta_2 wB + \delta_3 wC + \nu$는 사실상 아무 것도 분석하지 못합니다.

더미변수들을 가지고 분석모델을 구성할 때는 전체 분류에 속하는 더미변수들 중 하나를 제외하고 모델에 투입해야 합니다. 그 제외된 분류가 바로 전체 더미들에 대한 기준변수가 됩니다. 예를 들어보면, 종속변수에 대한 계절의 효과를 보고 싶다고 합시다. 만약 봄, 여름, 가을, 겨울을 각각 더미로 측정할 경우 그 중 봄을 제외한 나머지 세 계절의 더미변수를 모델에 투입합니다. 이때, 여름/가을/겨울의 계수는 봄이라는 계절에 비교하여 해석할 수

[1] 구분하기 편하게 연속형 변수는 양적 변수(quatitative variables)로, 명목형과 순위형 변수는 질적 변수(qualitative variables)로 사용하겠습니다.

있습니다. 일단 수리적 모형으로 살펴보겠습니다. 이번에는 사계절이 아니라 앞서의 $w \in \{A, B, C\}$의 분류형 변수를 기준으로 보겠습니다.

$$z = \gamma + \delta_2 wB + \delta_3 wC + v$$
$$z = \gamma + \delta_1 wA + \delta_3 wC + v$$
$$z = \gamma + \delta_1 wA + \delta_2 wB + v$$

첫 번째 식의 경우 δ_2, δ_3는 wA에 비하여 wB와 wC가 종속변수에 미치는 효과를 보여줍니다. $\delta_2 < 0$이라면 우리는 wB가 wA에 비해 종속변수에 미치는 효과가 δ_2만큼 '덜' 하다는 의미입니다. 간략히 말하자면, 더미변수들의 효과는 항상 생략된 카테고리에 비교하여 해석되어야 합니다.

4.6.2 상호작용(Interactions)

일반적으로 우리가 다중선형회귀모델을 수립한다고 할 때, $y = \alpha + \beta x + \gamma z + \nu$라고 할 수 있을 것입니다. 이때, y에 대한 x의 효과는 무엇으로 알 수 있을까요? $\partial y / \partial x$겠죠? 그리고 위의 식에서는 β일겁니다. 그런데 만약 $\partial y / \partial x$가 $beta + \gamma z$라면 어떻게 될까요? $\partial y / \partial x = \beta + \gamma z$인 경우를 상정해보겠습니다. 이 경우 β의 의미는 무엇일까요?

자, 상호작용을 보다 명시적으로 보여주는 회귀모델을 만들어보겠습니다.

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

이 회귀모델에서 β는 무엇을 의미할까요? 상호작용이란 y에 대한 x의 효과가 본질적으로 z라는 변수와 밀접하게 연관되어 있는 경우를 말합니다. 따라서 $z = 0$인 경우를 제외하면, x에 대한 계수값은 결코 그 자체로 그대로 해석될 수 없습니다. 왜냐하면 x의 계수값은 z의 값에 따라 조건적으로 변화할테니까요. 상호작용에 대한 이론적인 검토는 다음 장에서 이어서 하도록 하고, 여기서는 기본적인 내용들을 간단한 R 예제와 함께 살펴보도록 하겠습니다. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + u$의 형태를 취하는 회귀모델이 있다고 해보겠습니다.

사용할 데이터는 언제나와 같이 `Quality of Government`에서 가져오며 2016년 데이터입니다.

```
library(ezpickr)
library(tidyverse)
QOG <- pick(file = "http://www.qogdata.pol.gu.se/data/qog_bas_ts_jan19.dta")
QOG.s <-QOG %>%
  select(ccode, cname, year, wdi_agedr, wdi_trade, wdi_gdpcapcon2010) %>%
  dplyr::filter(year==2016) %>% drop_na()
```

`QOG` 데이터셋에서 국가코드, 국가명, 연도, 노령화 지수, 무역 개방성, 1인당 GDP에 해당하는 변수들을 따로 선별하여 서브셋을 만들고, 결측치를 제외하였습니다. 그리고 1인당 GDP가 무역 개방성, 무역 개방성의 제곱항, 그리고 노령화 지수와 각각 관계를 맺고 있다는 선형회귀모델을 구축하였습니다.

```
model <- lm(wdi_gdpcapcon2010 ~
              wdi_trade + I(wdi_trade^2) + wdi_agedr,
            data=QOG.s)
model %>% broom::tidy() %>%
  mutate_if(is.numeric, ~ round(., 3)) %>% knitr::kable()
```

term	estimate	std.error	statistic	p.value
(Intercept)	37929.275	6954.839	5.454	0.000
wdi_trade	-126.946	66.146	-1.919	0.057
I(wdi_trade^2)	0.806	0.195	4.135	0.000
wdi_agedr	-357.573	76.342	-4.684	0.000

구체적으로 위의 모델은 2010년도 미 달러 고정으로 측정된 2016년도의 1인당 GDP가 각 국가의 무역 개방성과 노령화 지수와 관계가 있다는 것을 보여주고 있습니다. 모델에서 무역 개방성은 국내 총생산에서 재화 및 서비스의 수출입의 총합이 차지하는 비율로 측정되었으며, 노령화 지수는 노동가능인구 대비 65세 이상 인구의 비율을 의미합니다. 이 회귀모델은 다중회귀모델의 형태를 취하고 있으므로($y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + u$), 우리는 y(1인당 GDP)의 변화량이 x(무역 개방성)와 z(노령화지수)에 의해 설명된다고 진술할 수 있습니다.

그러나 이 모델은 동시에 x와 x^2을 포함하고 있습니다. 이는 x의 y에 대한 한계효과(marginal effect)가 β_1 뿐 아니라 β_2에 의해서도 영향을 받는다는 것을 의미합니다. x를 기준으로 편미분을 해보면, $\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$가 되기 때문입니다. 정리하자면, 이 모델은 y의 변화량이 x와 z에 의해 설명될 수는 있지만, x와는 그 관계가 선형적일 것이라고 볼 수는 없다는 것을 의미합니다. 그러면 이 모델의 각각의 계수들(coefficients)을 해석해보겠습니다.

- 상수항(절편; intercept): $\hat{\beta}_0$는 PRF의 β_0의 추정치이자 x와 z 모두가 0일 경우의 y의 값입니다.
- 기울기(slopes)
  - $\hat{\beta}_1$는 PRF의 β_1의 추정치이자 x^2와 z로 설명될 수 있는 변화량을 제외한 y와 x 간의 관계를 보여줍니다. $\hat{\beta}_1$의 표준오차는 PRF를 추정하기 위해 우리가 수없이 표본들을 뽑아 PRF에 대응하는 SRF를 만들었을 때, 표본의 차이로 인해 각 SRF에서 나타날 $\hat{\beta}_1^i$들의 표집분포의 표준편차를 의미합니다.
  - $\hat{\beta}_2$는 PRF의 β_2에 대한 추정치를 의미하며, 순수하게 x와 z에 의해 설명되는 y의 변화량을 제외하고 x^2로 설명되는 y의 변화량에 대한 관계를 보여줍니다. 마찬가지로 그 표준오차는 PRF를 추정하기 위한 SRF에서 도출되는 $\hat{\beta}_2^i$의 표집분포의 표준편차를 보여줍니다.
  - $\hat{\beta}_3$는 x^2과 x로 설명되는 y의 변화량을 제외하고 y와 z 간의 관계를 보여주며, 그 표준오차는 $\hat{\beta}_3^i$의 표집분포의 표준편차라고 할 수 있습니다.

앞서 말했다시피 $\hat{\beta}_0$는 x와 z가 0일 경우의 y 값, 절편을 의미합니다. 이 값은 고정되어 있으므로 우리는 상수항이라고도 합니다. 이론적으로 변수인 y는 오직 다른 변수로만 설명할 수 있습니다. 따라서 우리는 기울기들에 좀 더 초점을 맞출 필요가 있습니다.

그런데 이 모델에서 우리는 $\hat{\beta}_1$, $\hat{\beta}_2$, 그리고 $\hat{\beta}_3$를 직접적으로 비교할 수 없습니다. 왜냐하면 $\hat{\beta}_3$는 y와 z 간 관계를 선형으로 상정하고 있지만 y와 x는 비선형성을 보여주고 있고, 특히 x^2과 y의 관계를 보여주는 계수는 그 자체로 설명될 수 없고 β_1에 의존적인 값이기 때문입니다. 따라서 이 모델이 다중선형회귀모델의 형태를 취하고 있다고 하더라도 $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$의 세 계수들은 직접적으로 비교되기는 어렵습니다. 또한 $\hat{\beta}_1$, $\hat{\beta}_2$는 x로 다시 써볼 수 있는데, 다시 한 번 말하지만 $\hat{\beta}_1$ 또는 $\hat{\beta}_2$는 단독으로는 의미가 없습니다.

상호작용항이 흥미로운 이유는 무엇일까요? 우리가 이차항(quadratic term)을 포함시키고, $\hat{\beta}_1$과 $\hat{\beta}_2$를 구했다고 할 때, 단적으로 말하면 $\hat{\beta}_1$과 $\hat{\beta}_2$ 그 자체의 값에는 크게 신경쓰실 필요가 없습니다. 만약 단순선형회귀모델이었다거나 다중선형회귀모델이었다면 $\hat{\beta}_1$과 $\hat{\beta}_2$가 x의 한 단위 변화와 관계된 y의 변화를 체계적으로, 일관되게 보여줄 것으로 기대하기 때문에 관심을 가질 필요가 있습니다.

그러나 이차항, 혹은 상호작용항을 모델에 포함한다면, x에 따른 y의 변화는 $\frac{\partial y}{\partial x} \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$로 나타낼 수 있고, 이는 x의 y에 대한 한계효과가 "변할 수도 있고," "비선형적일 수도 있다는 것"을 의미합니다. 따라서 $\hat{\beta}_1$과 $\hat{\beta}_2$ 그 자체는 x와 y의 관계에 대해서 거의 말해주는 것이 없습니다. $\hat{\beta}_1, \hat{\beta}_2$의 값과 그 크기보다는 x와 y의 비선형적 관계의 방향성을 보여줄 수도 있는 부호에 관심을 가지는 것이 더 나을 것입니다. 또한, 우리는 x에 대한 서로 다른 값들 간 한계 효과의 차이에 초점을 맞춰야 합니다. 예를 들어, x의 최소값과 최대값 간 한계효과를 비교함으로써 우리는 y의 변화 양상을 포착할 수 있기 때문입니다.

4.7 간단한 정리

4.7.1 개념들의 유기적인 연결

이 통계적인 논리들은 개별적인 것같지만 모두 유기적으로 연결되어 있습니다. 한 가지 예제를 통해 앞서 살펴보았던 내용들을 간단히 정리하는 시간을 가져보고자 합니다. 여러분들이 야구 경기가 시작하기 전에 그 앞에 위치한 식당의 맥주 소비량이 암표 가격(black market tickets prices)에 미치는 영향이 있는지를 연구해본다고 합시다. 그리고 어떤 정보상이 이에 관한 다양한 데이터들을 판매하는데 단 한 가지만 살 수 있습니다.

첫째, 아마도 "영향"을 연구하고자 하기 때문에 우리는 주로 회귀모델에서 기울기 계수값이 얼마나 큰지에 관심을 가질 것입니다. 그렇다면 문제는 정보상에게 어떤 데이터셋을 구매하는 것이 최선인가 하는 것입니다. + 큰 기울기 계수값을 갖기 위해서는 얼마나 많은 설명변수들이 서로를 설명하는지(covary), 그리고 얼마나 각 설명변수가 고유한 종속변수에 대한 설명력을 가지는지를 알아야 합니다. 아무리 설명변수가 많더라도 설명변수들끼리 공분산이 크다면, 정작 종속변수를 설명하는 데 중첩되어 개별 설명변수가 큰 계수값을 가지기 어렵습니다.

둘째, 만약 계수값의 크기보다 작은 표준오차를 더 신경쓴다면 어떤 데이터가 필요할까요?

- 일단, 표준오차가 무엇인지에 대해서 다시 한 번 생각해봅시다.
    - 이론적으로 우리는 관심을 가지고 있는 하나의 모집단으로부터 무한한 수의 표본들을 추출해낼 수 있습니다.
    - 우리가 관심을 가지고 있는 것은 모집단 수준에서의 계수들, PRF의 계수들이지만 실제로 모집단은 관측할 수 없기에 우리는 관측가능한 표본들을 가지고 PRF에 대응하는 SRF를 구성해 PRF의 계수들을 추론하게 됩니다.
    - 즉, 표본에 따라서 SRF에 따라 도출된 표본 통계치들은 다소 다르게 나타날 수 있습니다. 대표적으로 표집 방법 등과 같은 이유로 추출된 표본들이 완전히 동일할 가능성이 매우 낮기 때문입니다. 따라서 PRF는 특정한 값으로 정해져 있지만, 우리는 그것을 모르고 우리가 구한 SRF의 계수들이 그 PRF의 계수들, 모수(parameters)를 중심으로 분포하고 있다고 생각하게 됩니다.
    - 이때, 서로 다른 표본들로부터 각기 얻은 일련의 계수들(a set of coefficients from different samples)이 분포를 이룬다고 할 때, 그 분포의 표준편차가 표준오차입니다.
    - 표준오차는 우리의 SRF 계수값들이 실제 진실된 PRF의 모수값과 평균적으로 얼마나 떨어져있는지를 보여줍니다.
- 이 경우에는 표본의 크기에 대해 물어볼 필요가 있습니다. 표본의 크기가 커질수록 표준오차는 필연적으로 작아집니다. 또한 설명변수에 대한 전체 표본의 변화량(total sample variations)을 확인해보아야 합니다. x의 총 변화량이 커질수록, PRF의 x에 대한 β_1의 추정치 $\widehat{\beta_1}$의 표준오차는 더 작아지게 됩니다.

셋째, 누군가가 "다중공선성(mulicollinearity)은 항상 나쁘다"라고 말했다고 합시다. 과연 그럴까요? 앞서 회귀모델의 가우스-마르코프 가정 중 우리는 "완벽한 다중공선성이 없어야 한다"는 내용이 있다는 것은 이미 알고 있습니다. 그렇다면, 우리는 다중공선성을 완벽하게 피할 수 있을까요?

- 다중공선성은 가급적 적을 수록 좋겠지만, 항상 나쁜 것은 아닙니다. Wooldridge(2016: 84)에 따르면 다중공선성은 "둘 이상의 설명변수들 간 높은(하지만 완벽하지는 않은) 상관관계"라고 정의됩니다. 만약 x_1과 x_2가 매우 상관되어 있다면, 이는 y를 설명하기 위한 x_1의 고유한 변량과 x_2의 고유한 변량이 작을 것이라는 점을 시사합니다.
- 보다 기술적으로 $\widehat{\beta_j}$의 분산에 대해 생각해보겠습니다. 이후로는 $VAR(\widehat{\beta_j})$라고 하겠습니다. $VAR(\widehat{\beta_j}$는 $\sigma^2/[SST_j(1-R_j^2)]$로 나타낼 수 있습니다.
    - 이때 분자는 예측변수들의 영향력을 제외한 y의 변량입니다. 따라서, 이는 오차의 분산이라고 할 수 있습니다.
    - 분모는 다른 예측변수들의 변량을 포함하지 않은 순수한 각 설명변수의 고유한 분산입니다.
- 그러나 다중공선성은 항상 나쁘다고 하기에는 어려운 것이 우리는 어디까지나 예측변수들을 이론적 배경에 입각하여 선택하기 때문입니다. 만약 서로 다른 예측변수에 대한 구성개념들이 서로 다른 현상들을 차별적으로 보여준다고 한다면, 우리는 그 변수들이 매우 상관되어 있다고 하더라도 그것들을 제외하기 위해서는 타당한 이유를 갖추어야 합니다.
- 수적 공변(numerical covariation) 때문에 변수들 간 매우 높은 상관관계를 가질 수 있습니다만 그것이 반드시 나쁘다고 할 수는 없습니다. 단지 그러한 다중공선성이 높은 변수들을 포함한 모델이 설명력에 있어 "덜 유용하다"고 표현할 수 있을 따름입니다.
- Goldberger는 이 다중공선성의 문제를 "과소표본크기(micronumerosity)"라는 개념으로 설명하고 있습니다.

- 우리가 다중공선성을 설명변수들 간 높은 상관성으로 정의한다고 할 때, 그 결과로 우리는 더 큰 표준오차를 가지게 되어 결과적으로 추정치의 편의(bias)를 의심해볼 수 있습니다.

- 그러나 Goldberger는 예측변수들이 매우 상관되어 있다면 그 표준오차는 반드시 크게 나타날 것이며, 큰 표준오차는 모델의 변수들 간의 관계가 매우 불확실하다는 것을 의미한다고 주장합니다.

- 즉, 설명변수들 간 상관계수가 높아질수록 우리는 변수들의 변화가 종속변수와 관계된 것인지 아닌지를 구별하기가 어려워집니다. 따라서 Goldberger는 과소표본크기라는 개념을 통해 이 문제가 "작은 표본"에 따른 것으로 봐야 한다고 봅니다.

- 과소표본크기의 맥락에서 보자면, 우리가 충분한 크기의 데이터를 가지지 않는다면 표준오차가 더 커질 것입니다. 예측변수들의 높은 상관관계는 각 예측변수들의 고유한 변량이 매우 작다는 것을 의미합니다. 이는 곧 종속변수를 설명할 변량—정보가 부족하다는 것과 상통합니다.

4.7.2 변수에 대한 심화

우리가 $y_i = \delta_0 + \delta_1 x_i$인 모델을 가지고, 이때 $i \leq n$이라고 합시다. 그리고 $a = \frac{1}{n}\sum_i x_i$이고 $b = 2sd(x)$라고 하겠습니다. 이때, $y_i = \lambda_0 + \lambda_1 \frac{x_i - a}{b}$를 어떻게 추정할 수 있을까요? Gelman (2008)[2]의 내용에서 표준오차에 대해 조금 더 자세히 이해하기 위해 가져와 본 내용입니다.

단순하게 보겠습니다. 만약 $b = 2sd(x)$가 아니라 $b = sd(x)$였다면, 두 번째 수식에서 $\frac{x_i - a}{b}$는 각 관측치들에서 평균을 제외하고 그것을 표준편차로 나누는, 일종의 표준화(standardization) 작업이라고 이해할 수 있을 것입니다. 표준화 작업은 우리로 하여금 변수의 측정척도의 영향에서 자유롭게 합니다. 그런데 만약 $b = 2sd(x)$라고 한다면, 이 작업이 근본적으로 무언가 차이가 있을까요? 왜 우리는 꼭 $sd(x)$로 표준화를 해줘야만 하는 것일까요?

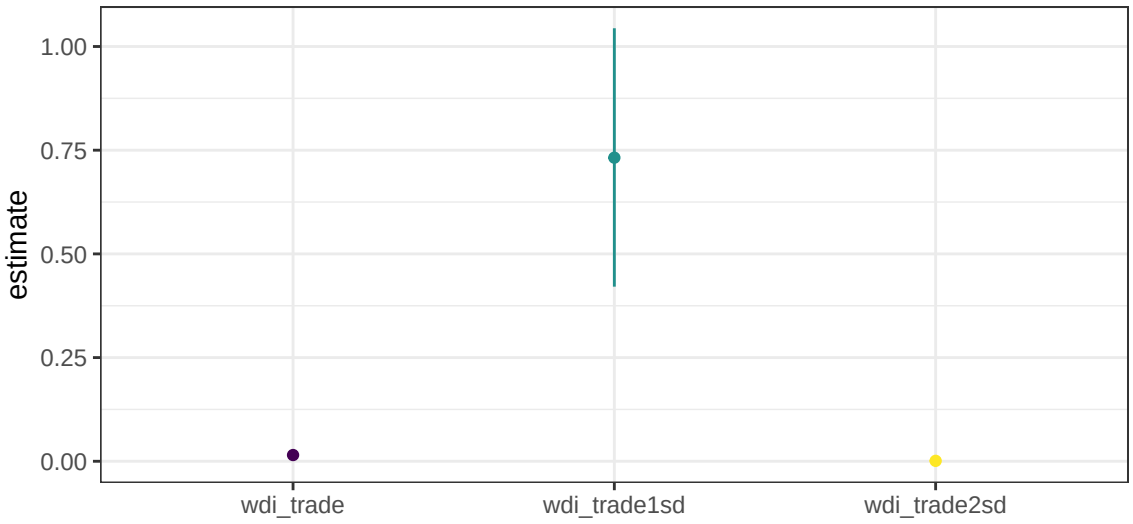
$y = \delta_0 + \delta_1 x_i$는 단순회귀모델입니다. 이때 δ_0는 절편값이고 δ_1은 모델의 회귀계수, 즉 x_i가 y와 맺는 관계 양상을 보여주는 것이죠. 그런데 우리가 두 번째로 수립한 수식에서 a는 x의 평균입니다. b는 x의 표준편차에 2를 곱한 것이죠. 따라서 a, b로 되어 있는 이 식을 x와 관련하여 다시 구성하여주면 다음과 같습니다: $y = \lambda_0 + \lambda_1(\frac{x_i - \bar{x}}{2 \times \sigma_x})$. 앞서 말했다시피 일종의 표준화 작업입니다. 이론적으로 $sd(x)$와 $2sd(x)$의 차이를 논의하기 전에 실제 데이터를 가지고 어떠한 편화가 있는지를 살펴보겠습니다. 이미 만들어놓은 `QOG.s` 데이터셋의 1인당 GDP와 무역 개방성 변수를 이용해보겠습니다.[3]

```
QOG.s %>% select(wdi_trade) %>% mutate(
  wdi_trade = wdi_trade,
  wdi_trade1sd = (wdi_trade - mean(wdi_trade))/sd(wdi_trade),
  wdi_trade2sd = (wdi_trade - mean(wdi_trade))/2*sd(wdi_trade)
) %>% summary() %>% knitr::kable()
```

	wdi_trade	wdi_trade1sd	wdi_trade2sd
	Min. : 20.72	Min. :-1.3093	Min. :-1616.3
	1st Qu.: 56.10	1st Qu.:-0.5973	1st Qu.: -737.3
	Median : 77.26	Median :-0.1714	Median : -211.6
	Mean : 85.78	Mean : 0.0000	Mean : 0.0
	3rd Qu.:102.75	3rd Qu.: 0.3415	3rd Qu.: 421.5
	Max. :407.43	Max. : 6.4735	Max. : 7991.1

[2] Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." Statistics in Medicine 27: 2865-73.

[3] 여기서 염두에 두어야할 것은 본래 `QOG.s`는 1인당 GDP, 무역 개방성, 그리고 노령화 지수를 대상으로 서브셋으로 만든 후에 결측치를 제거했기 때문에 노령화 지수가 결측치일 경우, 그에 해당하는 id의 1인당 GDP와 무역 개방성도 제외되었을 수 있습니다. 편의상 이전에 사용한 서브셋을 다시 사용하는 것이므로 이 점을 감안해야 합니다.



그럼 이번에는 세 변수 각각으로 종속변수를 선형회귀분석으로 추정했을 때의 결과를 비교해보도록 하겠습니다. 일단 이 예제를 진행하는 데 있어서 기술적으로(technically) 조작한(manipulated) 변수들의 내용은 다음과 같습니다.

- 첫째, 1인당 GDP는 달러 기준이므로 단위가 너무 커서 가시적으로 보여주기가 어렵기 때문에 로그를 취한 값을 사용했습니다. 무역 개방성은 비율이기 때문에 단위차이가 너무 크면 표준화한 결과와 원변수의 결과를 한 그래프 안에서 보여주기 어려울 테니까요.
- 둘째, 무역 개방성은 원변수(`Original`), 1표준편차로 표준화한 값(`One_std`), 그리고 2표준편차로 표준화한 값(`Two_std`)로 조작하였습니다.
- 절편값에는 크게 관심이 없으므로 제외하고 무역 개방성과 1인당 GDP 간의 선형회귀모델 결과만 보여드리도록 하겠습니다.

간단하게 정리하자면 예측변수의 통계적 유의성 여부, 방향성은 표준화 여부를 떠나서 같습니다. 그리고 여기서는 보여드리지 않았지만 R^2도 동일합니다. 그 얘기는 세 모델 모두 동일한 y의 변동량의 일부를 설명하고 있다는 것이겠죠? 그러나 $\hat{\beta}_x, \hat{\beta^{\sigma}_x}, \hat{\beta^{2\sigma}_x}$와 $se_{\hat{\beta}_x}, se_{\hat{\beta^{\sigma}_x}}, se_{\hat{\beta^{2\sigma}_x}}$는 서로 다릅니다.

만약 우리가 변수를 적절하게 표준화한다면, x와 y의 관계는 왜곡되지 않게 나타날 것입니다. 또한 변수들은 변수들의 평균에 영향을 받지만 동시에 변수들의 분산—표준편차에도 영향을 받습니다. 왜냐하면 회귀모델은 통계적 기법으로 얘기하자면 이른바 "평균 차이"를 검정하는 방법이기 때문입니다.[4] 그러나 변수를 표준화한 다음에는 x의 한 단위 변화라고 해석할 때, 그 한 단위가 무엇을 의미하는지 고민해볼 필요가 있습니다.

일반적인 표준화에 대한 함의가 위의 논의였다면, 다음은 Gelman (2008)의 제안대로 $2sd(x)$를 이용해 표준화한 결과가 무슨 의미가 있는지 살펴보겠습니다. 일반 표준화와 비교해서 무슨 차이가 있나요? 보시면 그래프에서 원변수의 계수의 양상과 $2sd(x)$ 표준화된 결과의 모습이 유사한 것을 확인할 수 있습니다. 즉, 측정척도의 영향에서 자유로우면서도 상대적으로 원변수의 경향성과 비슷한 모습을 보여줌으로써 우리는 단순 표준화 변수를 통한 결과보다 직관적인 이해가 가능하다고 할 수 있겠습니다.

[4] 어떤 변수의 영향력 하에서 종속변수의 평균이 그 변수의 영향력 외에 있을 때의 평균과 차이가 있는지 여부, 이것은 영가설 기각 논리의 배경이기도 합니다.