Chapter 6

Statistical Significance and Confidence Intervals

기초통계 파트와 현재 이 챕터까지 따라오셨다면, 회귀분석에 있어서 BLUE라는 것이 무엇을 의미하는지, 그리고 표집분포(sampling distribution)이 무엇인지, 적절하지 않은 변수들을 모델에 포함하였을 때 어떤 결과를 얻게 되는지와 적절한 변수를 모델에 포함시키지 않았을 때 생기는 문제 등에 대해서는 익숙하시리라 생각합니다.

바로 이전 챕터를 통해 이제는 $y = \alpha + \beta_1 x + \beta_2 z + \beta_3 xz + \epsilon$이라는 상호작용이 포함된 모델을 해석하실 수 있을 것이며, 나아가 $y = \alpha + \beta_1 I(x = 1) + \beta_2 I(x = 2) + \epsilon$, with $x \in \{0, 1, 2\}$와 같이 x가 이산형 변수(discrete)일 때 모델에 어떻게 투입하고 그 결과를 해석하실 수 있을 것이라 생각합니다.

이 챕터에서는 통계적으로 유의미하다는 것의 의미와 그 기준, 그리고 모집단을 알 수 없기 때문에 표본을 통해 통계적 추론을 한다는 것이 실제로 어떠한 분석을 수반하는지 등을 중점적으로 살펴보고자 합니다.

6.1 영가설 유의성 검정 (Null Hypothesis Significance Testing; NHST)

우리가 생각하는 모집단에 대한 β_k가 사실(true)라고 가정해보겠습니다. β_k가 보여주고자 하는 모수를 w라고 하겠습니다.

$\beta_k = w$라는 가정 하에서, 우리는 관측가능한 표본들로부터 추출한 $\hat{\beta}_k$들로부터 나타나는 분포가 얼마나 그 w를 보여주는 β_k를 포함할 가능성을 가지는지를 생각해볼 수 있습니다. 나아가 $\beta_k = w$라고 가정한다면 우리는 사실상 알 수 없는 모수를 아는 것과 다름없게 됩니다. 왜냐하면 β_k의 표집분포의 기대값이 w를 제대로 보여준다면, 이후에는 충분한 수의 표집과정을 통해 β_k를 충분히 확보만 하면 되는 것이니까요.

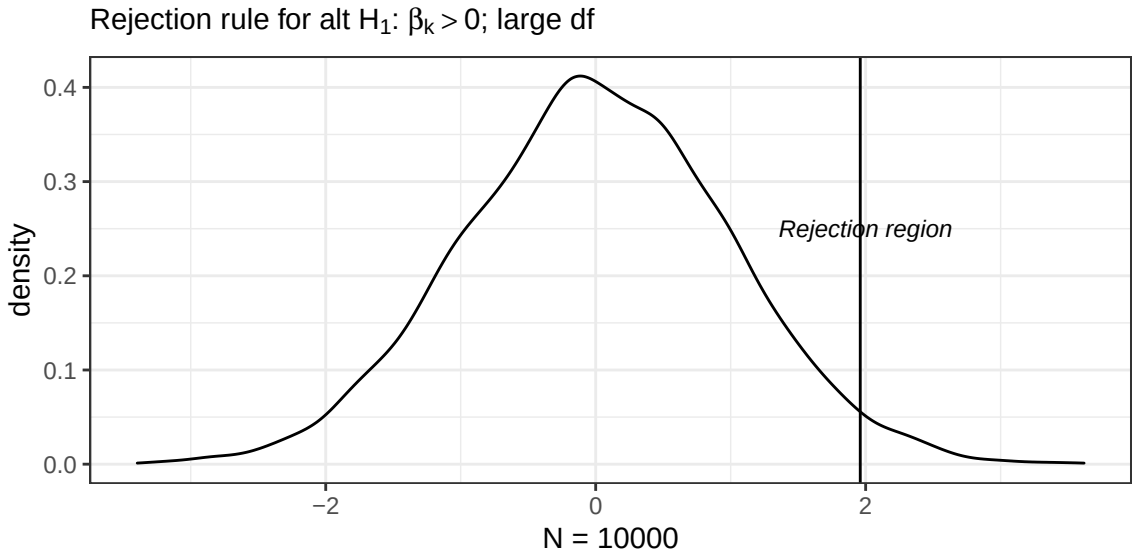
기술적으로 우리는 이것을 $(n - k - 1)$의 자유도를 가진 `Student t` 분포라고 합니다. 그리고 상당한 수의 관측치를 확보하게 된다면, 이 분포는 가우시안(Gaussian) 분포라고 할 수 있습니다. 아래와 같은 공식으로 나타낼 수 있겠네요.

$$\frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta})} \sim t_{n-k-1}$$

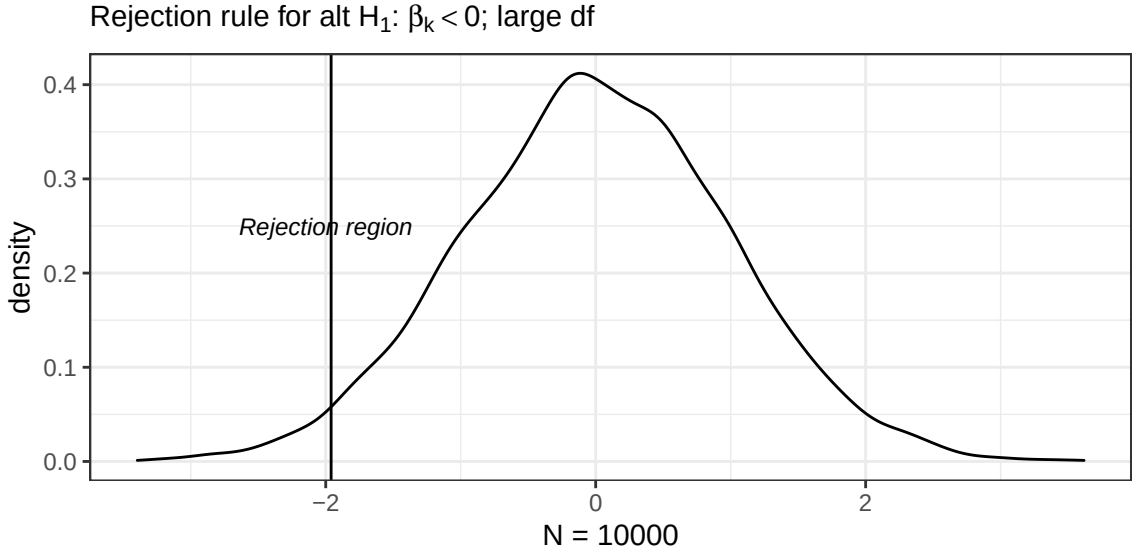
그렇다면 이제 영가설 유의성검정의 절차를 한 번 살펴보도록 하겠습니다. 영가설을 $H_0 : \beta_k = w$이라고 설정해보도록 하겠습니다. 그리고 이때 `t` 통계치는 $\frac{\hat{\beta}_k - w}{se(\hat{\beta}_k - w)} \sim t_{n-k-1}$로 나타낼 수 있겠네요. 그렇다면 여기서 문제입니다. β_k에 대한 영가설 혹은 연구가설을 기각하기 위해서 우리에게 필요한 `t` 통계치는 얼마일까요? 즉 $\hat{\beta}_k$은 얼마나 β_k로부터 떨어져 있어야할까요?

먼저 영가설 H_0 : $\beta_k = w$에 대해 그것과 비교하기 위한 대안가설(연구가설)을 H_1 : $\beta_k > w$이라고 해보겠습니다. 즉, 우리가 관심있는 것은 $\beta_k > w$이지만 이것을 직접적으로 검증하거나 혹은 확증할 수는 없기 때문에 $\beta_k \neq w$라는 것을 통해 영가설을 기각함으로써 영가설이 기각될 확률, 연구가설이 유의미할 '확률'을 구하게 되는 것입니다.

이것이 우리가 흔히 말하는 유의수준(significance level)입니다. 기술적으로 서술하자면 "모집단에서 추출한 100개의 표본 중에서 영가설이 사실일 경우를 기각하는 것을 몇 번 관측할 수 있느냐"라고 하는 것입니다. 만약 100개의 표본 중에서 12개의 표본이 영가설의 기대대로 $\beta_k = w$라는 것을 보여줬다면, 영가설이 사실일 확률은 `0.12`이 될 것입니다. 이때 우리는 이 영가설이 사실일 확률 `0.12`를 α라고 표현합니다. α의 값이 더 작을 수록 영가설을 기각하는 $\hat{\beta}_k$가 더 많다는 것을 의미합니다.



위의 플롯은 영가설(H_0)이 $\beta_k = 0$이라고 했을 때, 연구가설(H_1)이 $\beta_k > 0$인 경우, 관측치가 10,000개인 표본에서의 기각역(rejection region)을 보여주고 있습니다. 일단 전체 곡선 면적의 합은 1입니다. 당연하겠죠? 밀도함수로 나타낸 분포이니까 전체의 총합은 1입니다. 그리고 플롯에서 기각역이라고 나타난 선 우측의 면적의 총합은 0.05가 됩니다. 통상적으로 그것은 우리가 상정한 영가설대로 $\beta_k = 0$일 확률을 의미합니다.



이번에는 영가설(H_0)이 $\beta_k = 0$이라고 했을 때, 연구가설(H_1)이 $\beta_k < 0$인 경우, 관측치가 10,000개인 표본

에서의 기각역(rejection region)을 살펴보겠습니다. 마찬가지로 전체 곡선 면적의 합은 1입니다. 앞서의 플롯과 다른 점은 기각역의 위치입니다. 연구가설이 상정한 계수의 부호가 달라졌기 때문에 기각역의 위치도 달라진 것이죠. 따라서 이때는 선 좌측의 면적의 총합이 0.05가 되고, 영가설대로 $\beta_k = 0$일 확률을 의미합니다.

우리는 t 통계치가 우리가 상정한 100개의 표본 중에서 n개의 표본만이 영가설에 부합할 것이다라고 할 수 있는 일종의 결정적 기준값(critical value)보다 클 경우에 영가설을 기각할 수 있습니다. 이 결정적 기준값은 유의수준(α) 혹은 전체 확률에서 유의수준을 제한 값($1-\alpha$)을 Student t 혹은 정규누적밀도함수에 대입하면 구할 수 있습니다.

6.2 영가설의 기각 이후

영가설을 기각했다고 해보겠습니다. 그렇다면 우리는 그 기각 결과를 가지고 어떻게 연구가설에 대해 설명할 수 있을까요?

사실 α라고 하는 특정한 값을 사전에 미리 설정하고 그것을 선긋듯이 어떠한 결과를 결정하는 도구로 사용한다는 것은 문제의 소지가 있어 보입니다. 예를 들어 p 값이 0.051인 경우와 0,049일 때, 우리는 이들의 통계적 의미를 어떻게 이해해야할까요?

p 값은 주어진 t 통계치 하에서 영가설을 기각할 수 있는 가장 작은 α를 의미합니다. 공식을 통해서 살펴보자면, T가 우리가 얻을 수 있는 모든 가능한 검정 통계량(test statistics)라고 해보겠습니다. 예를 들어 Student t나 Gaussian이 있겠네요. 이때 만약 가설이 특정한 관계의 방향을 설정하지 않고 있다면, 단지 영가설 $\beta_k = 0$만을 기각하면 되기 때문에 이른바 양측검정을 상정한 연구가설을 설정하게 될 것입니다. 따라서 이때는 관계의 방향을 상정하지 않는 $Pr(|T| > |t|) = 2Pr(T > |t|)$가 성립하게 됩니다.

이제까지는 $\beta_k = 0$인 경우를 중심으로 살펴보았습니다. 왜냐하면 사실 회귀분석 같은 경우 영가설은 우리가 기대한 변수가 종속변수에 대해 '효과가 없을 것'을 주로 기대하기 때문입니다. 따라서 $\beta_k = 0$이 기각되었다는 것은 우리가 관심을 가지고 있는 변수가 종속변수에 대해 유의미한 영향이 있을 수 있다는 것을 시사합니다.

하지만 사실 영가설은 반드시 0, 효과가 없다는 것만을 기대할 필요는 없습니다. 예를 들어서 $H_0 : \beta_k = g$라는 특정한 값을 가지는 영가설에 대해서 $H_1 : \beta_k > g$와 같이 계수값이 그 특정한 값보다 클 경우를 상정할 수 있습니다. 이때 유의수준 α를 설정한다면, $t = \frac{\hat{\beta_k} - g}{se(\beta_k)}$가 되므로, 우리는 $t > c$이기만 하면 영가설을 기각할 수 있습니다.[1] 조금 더 풀어서 말하자면 일반적인 t 통계량은 다음과 같이 쓸 수 있습니다.

$$t = \frac{\text{추정치 - 가설로 기대하는 값}}{\text{추정치의 표준오차}}$$

이제 정치학 분야에서 실제로 사용할법한 데이터를 이용해서 위의 내용을 한 번 더 살펴보도록 하겠습니다. Quality of Government 데이터셋을 이용해서 $y = a_0 + \sum_k \beta_k x_k + e$의 형태를 갖는 선형회귀모델을 한 번 만들어 보겠습니다.

```
# 단순선형회귀모델과 다중선형회귀모델
model.simple <-
  lm(wdi_gdpcapcon2010 ~ wdi_trade, data=QOG.sub)
result1 <- summary(model.simple) %>% broom::tidy() %>%
  mutate(model = "SLR")
model.multiple <-
  lm(wdi_gdpcapcon2010 ~ wdi_trade + p_polity2 +
       wdi_pop1564, data=QOG.sub)
result2 <- summary(model.multiple) %>% broom::tidy() %>%
  mutate(model = "MLR")
results <- bind_rows(result1, result2) %>%
   mutate_if(is.numeric, round, 3)
```

[1] 여기서 c는 결정적 기준값, critical value를 의미합니다.

이렇게 만들어진 두 선형회귀모델을 가지고 영가설이 $\beta_k = 0$인 경우와 $\beta_k = c$인 경우를 한 번 살펴보도록 하겠습니다. 물론 이 경우에는 c가 합당한 비교 기준이라고 정당화할 수 있어야할 것입니다. 다중선형회귀모델을 기준으로 모델을 수식으로 표현한다면 아래와 같습니다.

$$\text{Economy} = \beta_1 \text{Trade} + \beta_2 \text{Level of Democracy} + \beta_3 \text{Working Population} + e.$$

```
results %>% knitr::kable()
```

term	estimate	std.error	statistic	p.value	model
(Intercept)	-390.277	2798.098	-0.139	0.889	SLR
wdi_trade	164.335	28.365	5.794	0.000	SLR
(Intercept)	-56094.546	12291.936	-4.564	0.000	MLR
wdi_trade	118.001	27.842	4.238	0.000	MLR
p_polity2	754.766	218.613	3.453	0.001	MLR
wdi_pop1564	889.742	201.305	4.420	0.000	MLR

여기서 x_1에 대한 계수, β_1을 중심으로 일단 영가설을 특정해보도록 하겠습니다.

- 먼저 $H_0 : \beta_1 = 0$이라고 할때, 민주주의 변수, 노동가능 인구비율이 통제되었을 때, GDP 대비 재화와 서비스의 수출 및 수입의 총합으로 측정된 무역개방성의 한 단위 증가는 2010년 고정 달러로 측정된 1인당 GDP, 즉 그 국가의 경제규모에 미치는 효과가 없다라는 것이 영가설임을 이해할 수 있습니다.
- 그 다음으로는 결정적 기준값으로 영가설을 수립해보도록 하겠습니다. 이번의 영가설은 $\beta_1 = 164.355$라고 하겠습니다. 즉, 무역 개방성의 한 단위 변화는 1인당 GDP로 측정된 경제규모가 163.48 증가하는 만큼의 효과를 가지고 있다는 것을 의미합니다.
- 여기서 결정적 기준값을 $\beta_1 = 164.355$라고 설정한 것은 단순선형회귀 모델에서의 무역 개방성의 계수 값입니다. 따라서 이 영가설을 통해 우리는 단순선형회귀모델에서 무역 개방성과 경제 규모 간의 관계가 다중선형회귀 모델에서 다른 통제변수들이 통제된 가운데 나타나는 무역 개방성과 경제 규모 간의 관계와 다른지, 그리고 그러한 차이가 통계적으로 유의미한지를 살펴볼 수 있습니다.

그렇다면 이번에는 우리가 관심을 가지고 있는, 연구가설에 대해 살펴보겠습니다. 효과의 방향(부호)를 생각할 필요가 없는 경우(양측, two-sided)와 방향도 고려해야 하는 경우(단측, one-sided) 모두를 특정해보도록 하겠습니다.

- 효과가 없음이 영가설일 때 ($H_0 : \beta_1 = 0$)
    1. 단측 연구가설 ($H_A : \beta_1 > 0$, 또는 $H_A : \beta_1 < 0$): 무역 개방성은 경제규모에 대해 긍정적 또는 부정적 효과를 가지고 있다. 이때의 연구가설은 긍정적 효과 또는 부정적 효과가 별개의 것이다.
    2. 양측 연구가설 ($H_A : \beta_1 \neq 0$): 무역 개방성은 경제규모에 대해 '효과가 있다.'
- 연구가설이 $H_A : \beta_1 = 164.355$일 경우
    1. 단측 연구가설 ($H_A : \beta_1 > 164.355$, 또는 $H_A : \beta_1 < 164.355$): 경제규모에 영향을 미칠 수 있는 다른 요인들을 고려하지 않을 때보다 고려했을 경우(통제했을 경우), 무역 개방성이 경제규모에 미치는 효과가 더 크다/작다.
    2. 양측 연구가설 ($H_A : \beta_1 \neq 164.355$): 경제규모에 영향을 미칠 수 있는 다른 요인들을 고려하지 않을 때와 고려했을 경우(통제했을 경우), 무역 개방성이 경제규모에 미치는 효과는 다르다(같지 않다).

이번에는 각각의 영가설과 연구가설에 대한 검정을 시행하고 그 결과를 제시 및 해석해보도록 하겠습니다.

```
# 바로 이전의 분석을 이용해보도록 하겠습니다.
# 단순선형회귀모델의 무역 개방성에 대한 계수와 표준오차, 자유도
b1.simple <- results[2,2]
se1.simple <- results[2,3]
simple.df <- model.simple$df

# 다중선형회귀모델의 무역 개방성에 대한 계수와 표준오차, 자유도
```

```
b1.multiple <- results[4,2]
se1.multiple <- results[4,3]
multiple.df <- model.multiple$df

# 첫 번째 연구가설 중 단측 가설입니다.
# HA: beta_1 > 0
pt(as.numeric((b1.simple-0)/se1.simple),
   simple.df, lower = FALSE)
```

```
## [1] 1.937543e-08
```

```
# 0.05보다 훨씬 작은 값을 확인할 수 있습니다.
# 즉, 영가설의 기대대로 표본에서 결과를 얻을 확률이 매우 낮다는 것이고
# 영가설을 기각하기에 충분한 경험적 근거를 확보했다고 할 수 있습니다.

# 양측 가설을 한 번 살펴보겠습니다. 방향을 고려할 필요가 없죠.
# HA: beta_1 != 0
2*pt(-abs(as.numeric((b1.simple-0)/se1.simple)), simple.df)
```

```
## [1] 3.875085e-08
```

```
# 방향을 고려할 필요가 없기 때문에 방향을 고려했을 때의
# 확률에 2를 곱해줍니다 (좌측 + 우측)

# 두 번째 연구가설(단순 vs. 다중) 중 단측 가설입니다.
# HA: beta_1_Multi > beta_1_Simple
pt(as.numeric((b1.multiple-b1.simple)/(se1.multiple-b1.simple)),
   multiple.df, lower = FALSE)
```

```
## [1] 0.3673704
```

```
# 0.05보다 훨씬 큰 값, 영가설은 기각되지 않습니다.
# 당연합니다. 다중회귀분석에서의 무역 개방성의 계수값과
# 단순회귀분석의 무역 개방성의 계수값은 일단 단순회귀분석의
# 계수값이 더 컸습니다. 따라서 위의 연구가설, 다중회귀분석의
# 계수값이 단순회귀분석의 그것보다 클 것이라는 기대는 기각되는 것이
# 우리의 상식에 부합합니다.

# HA: beta_1_Multi != beta_1_Simple
2*pt(-abs(as.numeric((b1.multiple-0)/(se1.multiple-b1.simple))), multiple.df)
```

```
## [1] 0.3886917
```

```
# 방향을 고려할 필요가 없기 때문에 방향을 고려했을 때의
# 확률에 2를 곱해줍니다 (좌측 + 우측). 마찬가지로 영가설 기각.
# 둘이 서로 같지 않다는 기대죠? 하지만 두 계수 값의 차이는 각 계수값이
# 서로 다른 표본에서 계산되어 가지게 될 표집분포에 따른 편차,
# 즉, 표준오차에 따라 고려했을 때, 계수의 차이가 유의미하기 보다는
# 표본의 차이에 따라 나타날 수 있는 차이라고 볼 수도 있습니다.
# 그 결과 영가설은 기각되지 않습니다.
# 두 효과는 통계적으로 차이가 있다고 말하기에는 영가설을
# 기각할 수 있는 충분한 경험적 근거를 확보하지 못한 것입니다.
```

6.2.1 β_k에 대한 가설검정?

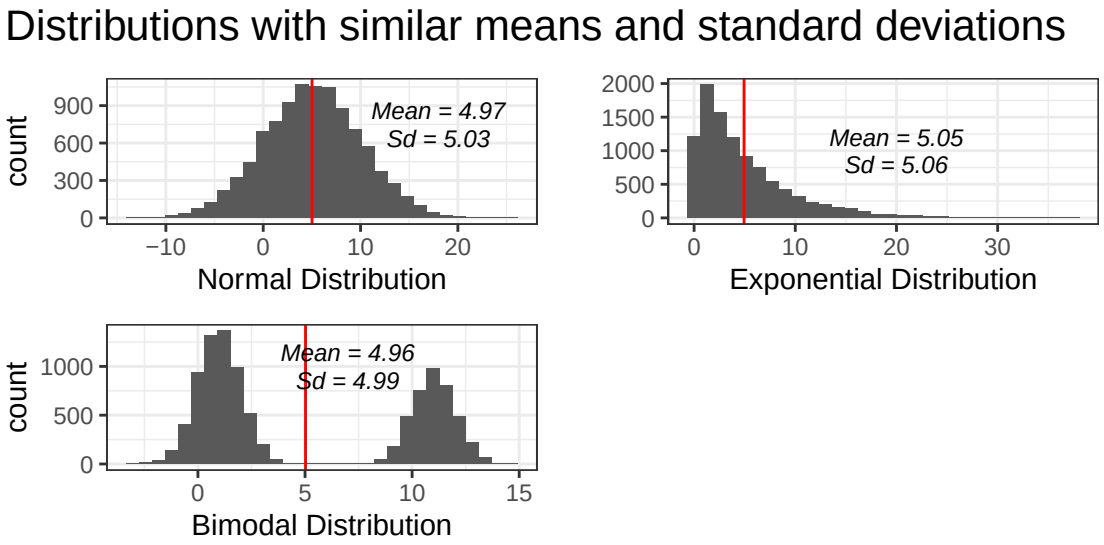
가설검정에 관련된 내용은 사실 영가설과 연구가설의 기각과 채택, 연역적 접근과 귀납적 접근에 대한 이론적 논의와 동시에 모집단과 표본에 대한 이해에서 시작됩니다. 만약 이 부분들에 대한 이해가 선행되지 않는다면, 같은 것을 말하고 있는 것 같지만 사실 다른 것을 의미하는 결과로 이어지기도 합니다.

예를 들어, 누군가가 β_k에 대한 가설검정을 해야한다고 합시다. 대충 들었을 때는 어떤 변수가 종속변수에 미치는 효과가 통계적으로 유의미한지, 혹은 어떠한 기준값에 비해 큰지 작은지 등과 같은 가설검정을 하고 싶다는 것으로 이해할 수 있을 것입니다. 하지만 저 '표현'에는 오류가 있습니다.

우리는 어떠한 관계를 보여주는 β가 모집단의 수준에서도 존재하는지 여부를 추론하기 위하여 가설을 수립합니다. 그러나 β_k는 단지 하나의 표본에서 추출한 하나의 통계치일 뿐입니다. 즉, 1번 표본으로부터 모집단의 β를 추정하기 위해 구한 것이 β_1이듯, k번의 표본으로부터 뽑은 통계치가 바로 β_k인 것입니다. 표집 과정의 본연적 한계로 인하여, 우리는 이론적으로 무수히 많은 표본들을 모집단으로부터 뽑을 수 있지만, 그 표본들로부터 얻을 수 있는 β_k들이 전부 동일하다고 기대할 수 없습니다. 따라서 하나의 표본에 대응하는 하나의 β_k에 대해 가설을 검증한다고 하는 표현은 우리가 실제로 관심을 가지고 있는 모수를 이해하는 데 있어서는 의미없다고 할 수 있습니다. 아 다르고 어 다른 것인데 이론적으로는 이렇게 큰 차이가 있습니다.

6.3 β_k와 $\hat{\beta}_k$에 대한 또 다른 접근

서로 다른 데이터를 보여주는 간단한 분포 세 개를 보도록 하겠습니다.



세 분포의 공통점과 차이점은 무엇일까요? 우선 세 분포는 평균(mean)과 표준편차(standard deviation)가 매우 비슷합니다. 하지만 실제 관측치들의 분포 양상은 매우 다르죠. 첫 번째 분포는 정규분포를 따른다면, 두 번째 분포는 지수분포, 마지막 분포는 크게 양극화된 양봉분포(bimodal distribution)의 형태를 띄고 있습니다.

단순히 데이터를 평균과 표준편차만으로 섣불리 이해할 경우 그 데이터를 이용한 통계적 추론에서 오류를 범할 수 있습니다. 따라서 우리는 '보다 구조화된' (more structured) 접근법을 취할 필요가 있습니다.

우선 가장 클래식한 선형회귀모델의 가정들을 다시 한 번 생각해보겠습니다. 총 다섯 개의 세부 가정으로 살펴볼 수 있는데, 순서대로 다중선형회귀분석(Multiple linear regression)에 대한 가정이라는 의미로 `MLR`라고 라벨링을 해보도록 하겠습니다. 어디까지나 여기서 클래식한 다중선형회귀분석의 가정들은 교차사례(cross-sectional) 자료에 적용가능합니다. 시계열(time-series)이 포함되면 또 고려해야할 문제들이 있기 때문인데, 이는 아마 고급 통계파트의 자료에서 다루게 될 것 같습니다.

- MLR1. 모집단에서의 모수들의 관계가 선형이다.

- MLR2. 무작위로 추출된 표본이다.
- MLR3. 설명변수들 간의 완벽한 다중공선성은 존재하지 않는다.
- MLR4. 설명변수와 오차는 독립적이다(Zero-conditional mean)
  - 주어진 설명변수라는 조건 하에서 오차항(u)의 기대값은 0이다.
  - 만약 이 가정이 성립되지 않는다면 우리는 모델 특정에 문제가 있었다고 생각할 수 있다. 예를 들면, 종속변수에 영향을 미치는 적절한 변수를 모델에 포함시키지 않아 오차항이 설명변수와 상관성을 가진다고 볼 수 있다.
  - 이 MLR4는 선형회귀분석의 추정치, 계수의 편향성 문제와 매우 밀접한 관계를 가진다.
- MLR5. (오차항의) 동분산성
  - 주어진 설명변수들에 대해 오차항의 조건 분포가 일정해야 한다.
  - 만약 오차항이 이분산성(heteroskedasticity)을 띈다고 하더라도 그것이 OLS 추정치가 편향되었다는 것을 의미하지는 않는다.
  - 다만 OLS 추정치의 '효율성'이 낮다는 것을 의미한다. 즉, OLS 중 `least` 의 조건 최소 분산(least variance)이 아니라는 것을 의미할 뿐이다.
- MLR1부터 MLR5까지가 충족되었을 때, 우리는 OLS의 결과를 `BLUE` (Best Linear Unbiased Estimates) 라고 한다.

자, 이 다섯 가정을 우리는 한 데 묶어서 Gauss-Markov 가정이라고 합니다. 이 클래식한 선형회귀모델의 가정을 조금 단순하게 묶어보자면 다음과 같이 표현할 수 있을 것 같습니다.

- $E(\hat{\beta}_k) = \beta_k$. 당연하겠죠? 만약 `BLUE`라면 우리가 표본들로부터 얻는 평균들의 평균이 모집단의 기대값과 일치할 거라는 가정 또한 충족될 것입니다.
- $Var(\hat{\beta}_k)$는 표본들로부터 얻은 평균들의 분산으로 어떠한 값을 가지게 될 것입니다. 뭐, 놀랄 건 아니죠. 모집단에서 표본을 추출하는 것은 본연적으로 일정한 오류를 내재하게 되어있기 때문에 평균들도 항상 모집단의 기대값과 같지는 않을 것이고, 그런 표본들의 평균들 간의 차이는 일종의 분산으로 나타날 것이니까요.

그런데 여기서 우리는 위의 두 발견으로부터 하나의 식을 도출해낼 수 있습니다. 만약 우리가 구한 OLS 추정치가 `BLUE`라면? 그래서 위의 두 조건이 충족된다면? 우리는 표본들로부터 구한 평균들이 어떠한 분포를 보일 것이라고 생각해볼 수 있습니다. 바로 모집단의 기대값을 보여줄 $E(\hat{\beta}_k)$를 중심으로 $Var(\hat{\beta}_k)$라는 분산을 가진 정규분포입니다.

$$\hat{\beta}_k \sim N(\hat{\beta}_k, Var(\hat{\beta}_k))$$

보통 통계학을 공부하다가 유의수준, 유의값, 신뢰구간 등을 배울 때 보면 순서가 연구가설/대안가설과 영가설의 관계, 가설 기각의 의미, 유의값의 의미, 그리고 잠정적인 1종오류와 2종오류의 문제, 점추정과 구간추정의 개념에서 살펴볼 수 있는 계수값과 신뢰구간의 관계와 의미 등으로 전개되는 것을 확인할 수 있습니다. 아마 `Lv.1.Stats_R`에서도 비슷한 접근법을 취했을 겁니다. 여기서는 조금 다른 방식으로 해당 주제들을 다루어보고자 합니다.

앞서 살펴보았던 $\hat{\beta}_k \sim N(\hat{\beta}_k, Var(\hat{\beta}_k))$가 성립한다고 할 때, 우리가 "알 수 있는 것"과 "알 수 없는 것"은 무엇일까요?

- 먼저 우리가 알 수 있는 것은 크게 세 가지라고 할 수 있습니다. 우리가 가지고 있는 예측변수이자 확률변수인 x가 어떠한 평균과 분산을 가진 정규분포를 따를 것이라는 것을 알 수 있습니다: x에 대한 분포 $\sim N(\mu, \tau^2)$: $\frac{1}{2\sqrt{\tau^2\pi}}\exp[-\frac{(x-\mu)^2}{2\tau^2}]$. 그리고 실제로 표본을 통해서 $\hat{\beta}_k$와 $Var\hat{\beta}_k$도 구할 수 있습니다.
- 반면에 모집단의 모수, β_k는 알 수 없습니다.

만약 우리가 β_k에 대한 가정을 세울 수 있고, 이론적 분석이 아니라 실제 경험적 데이터를 통해 그러한 가정이 충족되는지 아닌지를 확인할 수 있다면 어떻게 될까요?

6.4 신뢰구간(confidence interval)

영가설 유의성 검정, 즉 NHST에서 우리는 이미 α에 대해서 살펴보았습니다. 알 수 없는 모수 β_k에 대한 알고 있는 통계치 $\hat{\beta_k}$를 바탕으로 한 분포의 가정에서 우리는 다시 이 α로 돌아옵니다. α라는 개념의 바탕에는 "이론적으로 무수히 많은 표본"이 전제되어 있습니다. 적어도 이론적 수준에서 우리는 하나의 모집단으로부터 수많은 표본을 반복하여 추출할 수 있습니다. 그렇게 무수히 많은 표본들로부터 얻은 통계치의 분포를 표집분포(sampling distribution)이라고 할 수 있고, 그 표집분포를 바탕으로 우리는 얼마나 많은 표본이 영가설의 기대에 부합하는 통계치를 가지는지, 전체 표본에서 그러한 표본의 비율을 α라고 나타냅니다. 따라서 흔히 사용하는 유의수준 0.05라는 기준은 이런 맥락에서 보자면 "100번의 반복 추출된 표본 중에서 5개의 표본의 표본평균이 영가설의 기대를 포함하고 있을 경우"를 말한다고 할 수 있습니다. 그러면 α는 조금 더 직관적으로 이해할 수 있습니다.

- α는 표집분포에 있어서 일종의 꼬리확률(tail probability)이라고 생각할 수 있습니다.
- 신뢰구간은 우리가 일반적으로 모수가 그 사이에 존재할 것이라고 기대하는 구간입니다.
- 반복해서 추출한 표본들에 대해 95%의 신뢰구간은 100번 추출한 표본의 $\hat{\beta_k}$ 중에서 95개의 표본이 모수 β_k를 포함한다고 기대할 수 있습니다.

6.4.1 신뢰구간 구하기

표집분포로부터 우리는 $\frac{\hat{\beta_k}-\beta_k}{se(\hat{\beta})} \sim N(0,1)$이라는 것을 알 수 있습니다. 이건 t죠. 풀어서 말하자면 표본들을 통해 얻은 통계치들과 모수와의 차이를 표집 과정에서 나타날 수 있는 오차로 나누어준 값의 분포를 구할 수 있다는 겁니다. 일종의 표준화 작업을 해주었으니 분포는 평균 0을 갖는 정규분포로 수렴하게 될 것입니다. 여기서 평균이 0이라는 얘기는 어떤 표본에서 얻은 통계치가 완벽하게 모수의 값과 일치한다는 얘기겠죠?

95%의 신뢰구간을 양측꼬리확률로 생각해보겠습니다. 표집분포로 얘기해보자면 1000개의 표본 중 950개는 신뢰구간에 모수의 값을 포함하고 있지만 50개는 포함하지 않는 표본이라는 얘기일 것입니다. 그런데 이때 $\hat{\beta_k}$는 모수보다 매우 커서 신뢰구간에 모수의 값을 포함하지 않을 수도 있고, 모수보다 매우 작아서 그럴 수도 있습니다. 즉, 분포의 좌측과 우측 모두에서 모수를 포함하지 않을 확률을 더해서 5%라는 것입니다. 그렇다면 이를 분위(quantiles)로 보면 1000개의 표본에서 얻은 $\hat{\beta_k}$를 작은 값부터 큰 값까지 일렬로 줄을 세운다고 할 때, 제일값이 작은 통계치를 가진 표본부터 25번째로 작은 표본까지와 975번째로 작은 표본부터 마지막 표본까지가 이 5%에 해당할 것입니다. 그렇다면 기준점은? 25번째 표본에 해당하는 값과 975번째에 해당하는 값이겠죠(2.5th Quantile, 97.5th Quantile)? 이미 본 적이 있습니다. 정규분포에서 해당하는 확률을 가지는 t값은 $±$1.96입니다. 그렇다면 우리는 95% 신뢰구간을 이론적으로는 평균에서 $±1.96×$표준오차를 통해, 만약 1000개의 표본이 있다면 2.5분위와 97.5분위에 해당하는 값을 기준으로 삼을 수 있을 것입니다. 99%의 신뢰구간도 마찬가지일 것입니다. 정리하자면 신뢰구간은 이론적으로 표본평균과 표준오차를 알고 있다면 다음과 같이 구할 수 있을 것입니다.

95% 신뢰구간 : $[\hat{\beta_k}-1.96\times se(\hat{\beta_k}), \quad \hat{\beta_k}+1.96\times se(\hat{\beta_k})]$.99% 신뢰구간 : $[\hat{\beta_k}-2.58\times se(\hat{\beta_k}), \quad \hat{\beta_k}+2.58\times se(\hat{\beta_k})]$.

여기까지는 앞서 살펴보았던 NHST의 과정과 크게 다르지 않습니다. 그런데 주목해야할 것은 우리가 아까 β_k와 $\hat{\beta_k}$, 그리고 $Var(\hat{\beta_k})$ 간의 관계를 일종의 분포로 보여주었던 가정입니다.

1. 과연 실제로도 무수히 많은 표본을 뽑았을 때, 이론적 기대와 같은 분포가 나타날까?
2. 만약 β_k가 OLS의 가정인 Gaus-Markov 가정을 충족시켜 얻어낸 BLUE라고 한다면, 그래서 우리가 $\hat{\beta_k} \sim N(\hat{\beta_k}, Var(\hat{\beta_k}))$라고 할 수 있다면? 이러한 조건을 가진 분포에서 실제로 표집분포처럼 표본을 추출해서 이론적으로 설정된 신뢰구간이 아닌 정말로 2.5분위와 97.5분위에 해당하는 값을 통해 신뢰구간을 보여줄 수 있지 않을까요? 아니, 나아가 보다 구체적으로 $\hat{\beta_k}$의 분포를 직접적으로 보여줄 수 있지 않을까요?

이러한 기대를 가지고 다음 챕터에서는 최소자승법을 행렬적 기법으로 접근하는 내용과 비모수 부트스랩(Non parametric bootstrap)이라는 일종의 시뮬레이션에 관한 내용들을 정리해보도록 하겠습니다. 이제 이론만 다루는 파트는 거의 끝나고 실제로 R을 같이 돌리는 작업들을 해볼 수 있겠네요.