Chapter 7

Better Understanding for Linear Model

제6장에서 단순선형회귀모델(simple linear regression model)에 대한 기본적인 내용들을 주로 R코드와 통계치들을 중심으로 살펴보았다면, 이 장에서는 설명변수가 2개 이상인 다중선형회귀모델(multiple linear regression model)로 나아가기 이전에 짚고 넘어가야할 선형모델에 대한 이론적 배경들을 짚어보고자 합니다.

이론적인 측면에서 제4장과 제5장에서 이어진다고 할 수 있습니다. 구체적으로 이 장은 추정(estimation)-유의성 검정(significance tests)-선형회귀모델과 상관관계 순으로 앞선 장들에서 놓치고 넘어갈 수 있었던 이론적인 논의들을 다루어보고자 합니다.

7.1 추정(Estimation)

첫 번째 절에서는 추정의 문제를 살펴볼 것입니다. 통계방법을 이용하여 우리는 무엇을 추정하려고 하는 것인지, 그리고 추정에는 어떠한 종류가 있는지, 데이터에 따라 우리가 추정해야하기 위해 주목해야하는 것이 무엇인지 등을 다룹니다.

7.1.1 점추정과 구간추정(point estimation and interval estimation)

무엇인가를 추정한다는 것은 표본과 모집단의 관계로 설명할 수 있습니다. 현실세계 속에서 우리가 사회현상에 대한 모집단을 관측한다는 것은 불가능에 가깝습니다. 따라서 우리는 관측가능한 자료, 표본(sample) 데이터를 통하여 모집단의 모수(parameters)의 값을 "추정"하고자 합니다. 즉, 추정의 목적은 본질적으로는 표본을 통하여 모집단의 특성을 이해하는 것에 있습니다. 이 지점에서 제5장에서 살펴보았던 "추론"의 개념이 등장합니다. 모집단을 확보할 수 없으니, 모집단을 잘 대표할 것이라고 기대되는 그 일부, 표본을 통해 모집단이 이러이러할 것이다—라는 추론(inference)을 하게 되는 것이지요.

정량적 변수들을 이용해서 우리는 모집단의 평균(mean)을 추정할 수 있으며, 명목형 변수-분류형 변수를 이용해서는 각 부류가 전체 분류에서 차지하는 비율(proportion)을 추정할 수 있습니다.

통계학에서는 여러 가지 유사한 용어들이 나오고, 그것들을 구분하여 사용하는 것이 중요합니다. 예를 들어, 지금 언급하고 있는 추정(estimation)과 추정량(estimator)은 서로 다른 개념입니다.

- 추정량(estimator)이란 표본의 비율(sample proportion), 표본의 평균(saple mean)과 같은 모집단의 모수를 추정하기 위한 특정한 통계치(statistic)을 의미합니다.
- 이때 통계치(statistic)은 모집단의 모수(parameter)와 대칭적 관계로 이해할 수 있으며, 표본의 통계적 특성을 의미한다고 볼 수 있습니다.
- 한편, 추정값(estimate)은 표본의 특정한 값을 지칭합니다. 예를 들어, 표본의 비율(estimator)의 값(estimate)이 0.73이라는 식으로 표현할 수 있습니다.

모든 추정이 동일하지는 않습니다. 왜냐하면 우리는 모집단을 알지 못하여 표본을 가지고 모집단의 특성을 추정, 추론하기 때문에 우리의 추정에는 불확실성(uncertainty)이 존재하는 탓입니다. 여기서는 점추정량(point estimate)과 구간추정량(interval estimate)을 살펴보고자 합니다.

- 점추정량이란 모수의 값(parameter value)에 대해 최선의 추측이라고 할 수 있는 단 하나의 통계치(single statistic value)를 의미합니다. 즉, 표본을 통해 모집단의 평균이 0.5일 것이라고 말할 때, 우리는 그 0.5를 점추정량이라고 말할 수 있습니다.
- 반면, 구간추정량은 점추정량 둘러싼 구간을 의미하는데, 모수값을 포함할 것이라고 기대되는 "고정된 신뢰수준"(fixed confidence level)을 보여주는 구간[1]이라고 할 수 있습니다. 이 점에서 우리는 이 구간추정량에 대해 신뢰구간(confidence interval)이라고 합니다.

추정, 추정량, 추정값을 이해하는 것은 아마도 모집단의 모수를 제대로, 잘 추정하는 추정량을 얻기 위해서일 것입니다. 그렇다면 좋은 추정량은 무엇일까요? 크게 두 가지 기준을 생각해볼 수 있습니다.

첫째는 불편성(unbiasness)입니다. 이는 추정량의 표집분포가 모집단 값을 중심으로 분포되어 있어야 한다는 것을 의미합니다. 즉, 추정량은 표본을 바탕으로 한 것이기 때문에 모수를 추정하는 데 있어서 불확실성을 내표할수밖에 없지만, 그 불확실성이 모수를 중심으로 퍼져있는 것이라면 우리는 기대값—평균이나 비율 등을 이용하여 모수를 추정할 수 있습니다. 만약 추정량이 편향되어 있다면(biased), 평균적으로 모수를 과소추정(underestimate)하거나 과대추정(overestimate)한다면 우리는 그 결과를 신뢰하기 어려울 것입니다.

두 번째로는 효율성(efficiency)이 있습니다. 가능한 한 작은 표준오차를 통하여 추정할수록 우리는 그 추정량이 다른 추정량에 비하여 모수를 평균적으로 더 '가깝게' 추정하는 효율적인 추정량일 것이라고 기대하게 됩니다. 표준오차가 크다는 것은 그만큼 불확실하다는 것이고 우리가 모수값을 추정하는 데 믿기 어려운 결과라는 의미입니다. 그렇다면 우리는 그 불확실성을 보정하기 위하여 추가적인 조치들을 취해야할 것이고, 그 결과는 비효율적(inefficient)이라고 할 수 있습니다.

동시에 신뢰구간의 문제도 함께 생각해볼 필요가 있습니다. 신뢰구간은 앞서 언급했다시피 모수값을 포함하고 있을 것이라고 기대되는 수의 구간(interval of numbers)입니다. 그리고 그 구간을 계산해내는 방법이 모수를 포함할 것이라고 기대되는 확률을 우리는 신뢰수준(confidence level)이라고 정의합니다. 사회과학에서 대다수의 연구들은 0.95나 0.99와 같은 거의 1에 가까운 신뢰수준을 사용합니다.

신뢰구간은 대개 **점추정값** $\pm$ **오차한계(margin of error)**로 표현되며, 이때 오차한계란 점추정량의 표집분포(sampling distribution)가 어떻게 분포되어 있는지를 보여주는 것이라고 할 수 있습니다. 즉, 수없이 표본을 많이 뽑아봤을 때, 표집이라는 과정에서 본질적으로 불확실할수밖에 없는 점추정값이 어떻게 분포하느냐를 보여주는 것이죠. 따라서 오차한계는 점추정량에 대한 신뢰수준에 따라 달라지는데, 95% 신뢰수준에서 오차한계는 대략 점추정치로부터 $\pm 2\times$ 표준오차 정도라고 할 수 있습니다.

7.1.2 신뢰구간: 평균과 비율

데이터의 특성에 따라 우리는 서로 다른 추정량을 통해 모집단에 대해 추론합니다. 예를 들어, 정량적-연속형 변수일 경우에는 평균을 통해 그 자료를 대표하는 값을 보여줄 수 있지만, 명목형-분류형 변수일 경우에는 비율을 대표값이라고 할 수 있습니다. 따라서 신뢰구간을 구하는 것도 데이터에 따라서 다르게 보여줄 수 있습니다.

7.1.2.1 비율의 신뢰구간

명목형-분류형 변수일 경우, 우리는 비율에 대한 신뢰구간을 보여줄 수 있습니다. 표본의 비율은 $\hat{\pi}$로 나타내며, 이는 어떤 카테고리에 따른 존부(存否)를 보여주는 것이라고 할 수 있습니다. 예를 들어, 이항변수(binary variable)일 때, $\hat{\pi}$는 전체 관측치 중에서 $y = 1$가 몇 개 속해있는지를 보여주는 평균이라고 할 수 있습니다.

모집단의 비율은 즉, $P(1) = \pi$와 $P(0) = 1 - \pi$라는 확률분포의 평균 μ라고 할 수 있습니다. 이때, 이 확률분포의 표준편차는 $\sigma = \sqrt{\pi(1-\pi)}$로 나타낼 수 있으며, 표본 비율에 대한 표준오차는 $\sigma_{\hat{\pi}} = \sigma/\sqrt{n} = \sqrt{\pi(1-\pi)/n}$이라고 할 수 있습니다.

수리적으로 중심극한정리(Central Limit Theorem, CLT)에 따라서 표본 비율에 대한 표집분포는 무작위 표본의 수가 많으면 많아질수록 정규분포에 근사(approximation)하게 됩니다. 따라서 앞서의 신뢰수준 0.95의 확률에서 표본 비율 $\hat{\pi}$는 모집단의 비율 π를 둘러싼 1.96 $\times$ 표준오차의 구간에 존재한다고 말할 수 있습니다.

- 다르게 표현하면, 95%의 확률로 표본의 비율, $\hat{\pi}$는 모집단의 비율로부터 표집으로 인한 오차 사이의 구간 사이에 존재한다는 것입니다: $\pi - 1.96\sigma_{\hat{\pi}}$와 $\pi + 1.96\sigma_{\hat{\pi}}$.

[1] 유의성 검정 부분에서 한 번 더 짚고 넘어갈 것이지만, 여기에서 의미하는 신뢰수준이란 표본과 모집단, 그리고 표집(sampling)의 문제와 관련이 있습니다. 우리가 얻는 특정한 추정량들은 표집 방법에 따라 충분히 표본이 모집단에 대해 대표성을 가진다고 할 때, 신뢰할 수 있는 결과로 받아들여집니다. 그럼에도 불구하고, 모집단에서 표본을 뽑아낸다는 본연적 한계로 인하여 우리는 표본에서 얻어낸 결과가 모집단을 제대로 보여주지 못했을 결과를 감안해야만 합니다. 신뢰구간이란 이론적으로 총 몇 번의 표집과정을 되풀이했을 때, 그 중 얼마만큼의 잘못된 추정을 가질 확률을 의미합니다. 예를 들어, 100번의 표집을 거쳐 5번의 다른 결과를 얻었을 때, 우리는 95%의 확률로 그 결과값을 신뢰할 수 있다고 보는 것입니다.

- 따라서 선택된 표본에서 우리는 95%의 확률로 신뢰구간 $\hat{\pi} - 1.96\sigma_{\hat{\pi}}$와 $\hat{\pi} + 1.96\sigma_{\hat{\pi}}$이 모집단 비율 π를 포함하고 있을 것이라고 주장할 수 있습니다.

신뢰구간을 구할 때, 앞에서 표준오차를 $\sigma_{\hat{\pi}} = \sigma/\sqrt{n} = \sqrt{\pi(1-\pi)/n}$라는 공식으로 구할 수 있다고 했습니다만, 사실 이것은 이론적인 주장입니다. 왜냐하면 우리는 모집단의 표준편차, σ를 모르기 때문입니다. 때문에 실제로는 우리는 표본비율을 통하여 표집오차를 추정해냅니다.

$$se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

그러므로 95% 신뢰수준에서 표본비율의 신뢰구간은 $\hat{\pi} \pm 1.96(se)$라고 할 수 있습니다. 한 번 예제를 풀어볼까요?

18세부터 22세 사이의 미국인 중 몇 퍼센트가 "매우 행복함"이라고 응답했을까요? GSS 데이터를 이용해서 전체 164명 중에서 35명이 "매우 행복함"이라고 응답했다는 것을 확인했다고 합시다.

- 표본을 통해서 본 전체 18-22세 중 매우 행복함이라고 응답한 비율은 $\hat{\pi} = 35/164 \approx .213$입니다. 편의상 0.213이라고 하겠습니다.
- 그렇다면 표준오차는 $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$이므로, $\sqrt{0.213(0.787)/164}$라고 할 수 있습니다. 그 결과값이 0.032라고 합시다.
- 95% 신뢰구간은 $0.213 \pm 1.96(0.032)$가 되므로 0.213 ± 0.063이라고 할 수 있습니다. 오차한계가 0.063인 것입니다. 이는 구간, $(0.15, 0.28)$을 산출합니다.

결과적으로 우리는 "매우 행복한" 사람들이 모집단에서 차지할 비율이 95%의 신뢰수준에서 0.15와 0.28, 전체의 15%와 28% 사이에 위치할 것이라고 주장할 수 있습니다.

만약 99% 수준의 신뢰수준에서 신뢰구간을 찾고자 하면 어떻게 될까요? 신뢰구간은 오차한계에 따라 변화합니다. 오차한계는 표준오차를 신뢰수준의 기준값으로 곱해준 결과입니다. 신뢰수준이 변화하면, 기준값도 변화합니다. 99%의 신뢰수준은 양쪽 분포에서 0.5%씩을 차지하므로, 이때 z-score는 2.58입니다. 따라서 99% 수준의 신뢰구간은 위의 예제대로라면 $0.213 \pm 2.58(0.032) = 0.213 \pm 0.083$이므로, 구간은 $(0.13, 0.30)$이 됩니다. 즉, 신뢰수준이 높아질수록, 신뢰구간은 넓어집니다.

한편, 신뢰구간은 표본의 크기에도 영향을 받습니다. 위의 예제에서 표본의 수가 164개가 아니라 656개라고 생각해보겠습니다. 그러면 먼저 표준오차가 영향을 받습니다. $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$이므로, $\sqrt{0.213(0.787)/656} = 0.016$이라는 결과를 얻게 됩니다. 앞서 164개일 때(0.032)와 비교하면 표준오차가 작아진 것을 알 수 있습니다. 따라서 이때는 95% 신뢰수준에서 신뢰구간은 $0.213 \pm 1.96(0.016) = 0.213 \pm 0.031$로 $(0.18, 0.24)$가 됩니다. 즉, 표본의 규모가 클수록 신뢰구간은 작아지는 것을 확인할 수 있습니다.

정리하자면, 만약 우리가 정해진 규모(n)의 표본을 무작위로 반복해서 뽑고, 매번 95% 신뢰구간을 계산한다면, 장기적으로 보았을 때 95% 신뢰구간은 모집단의 비율을 그 구간 안에 포함하게 될 것이라고 기대할 수 있습니다.

- 비율에 대한 신뢰구간의 공식은 다음과 같습니다: $\hat{\pi} \pm z(se)$ with $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$

통계방법은 대규모 관측치를 통하여 표본비율에 대한 표집분포가 정규분포에 근사하여(CLT), 추정값에 대한 표준오차를 최대한 작게 만들 것을 요구합니다. 수리적으로는 분류형 변수에 있어서 최소 30개의 관측치가 존재하고, 그 중 최소 15개씩의 존부(1;0) 결과가 존재한다면 CLT에 따라서 추정할 수 있다고 봅니다. 만약 이러한 조건이 충족되지 않는다면, 표집분포는 왜곡될 수도 있습니다(skewed). 이는 표본비율이 모집단 비율(π)을 추정하는 데 쓸모없는 추정치가 될 수 있음을 의미하며, 표준오차도 모집단 수준의 표준오차를 추론하는 데 무용할 수 있다는 것을 의미합니다.

7.1.2.2 평균의 신뢰구간

이번에는 정량적-연속형 변수일 경우에 신뢰구간을 구하는 방법을 살펴보겠습니다. 관측치의 개수가 많은 무작위 표본에서, 표본의 평균은 모집단 평균 μ와 표준오차를 둘러싼 정규표집분포(normal sampling distribution)에 근사한다고 할 수 있습니다.

$$\sigma_{\bar{y}} = \sigma/\sqrt{n}$$

따라서, $P(\mu - 1.96\sigma_{\bar{y}} \leq \bar{y} \leq \mu + 1.96\sigma_{\bar{y}}) = 0.95$라고 정리할 수 있습니다. 이때 우리는 "95%의 확률로 표본평균이 (알 수 없는) 모집단 평균을 둘러싼 표준편차의 1.96배 구간 내에 위치한다"고 할 수 있습니다.

여기서의 문제는 우리가 표준오차를 모른다는 것입니다. σ도 모집단의 표준편차로, 모수이기 때문에 우리가 알지 못합니다. 따라서 우리는 σ를 알 수 없기 때문에 이를 표본 데이터로부터의 점추정값으로 대체해서 추정하게 됩니다. 즉,

$$se = s/\sqrt{n}$$

이 되는 것입니다. μ에 대한 95%의 신뢰구간에서 우리는 $\bar{y} \pm 1.96(se)$라고 주장하지만 실상 그것은 $\bar{y} \pm 1.96(\frac{s}{\sqrt{n}})$과 같습니다.

이러한 방법은 어디까지나 표본의 표준편차(s)가 모집단의 표준편차(σ)에 대해 좋은 추정값이라고 말할 수 있는 CLT의 근거인 "대규모 관측치"가 가정될 때 가능합니다. 만약 관측치의 개수가 적다면, σ를 s로 대체하는 것은 또 다른 오류를 불러올 수 있기 때문에, 신뢰구간이 충분한 너비를 가지지 못할 수 있습니다. 이때 우리는 z-score보다 조금 더 큰 값을 가지는 t-score를 기준치로 대신 사용해줌으로써 표본 규모에 따른 문제를 해결하고자 합니다.

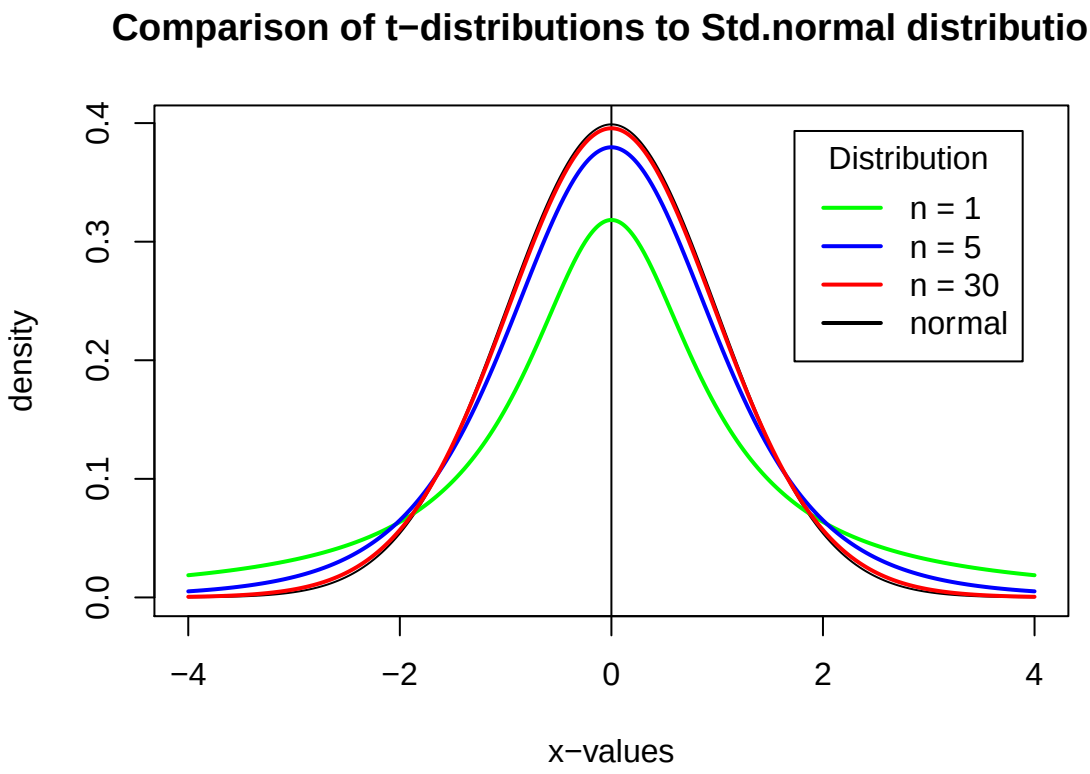
t-분포(`Student's t; the t distribution`)는 다음과 같은 특징을 가지는 분포입니다.[2]

- 0을 중심으로 종 형태(bell-shaped)를 취하는 분포
- 표준편차는 1보다 조금 큰 분포; 따라서 표준정규분포보다 양 꼬리가 조금은 더 두꺼운 모습을 보임.
- 정확한 형태는 자유도(degree of freedom, df)를 따르는데, 평균을 추론할 때 df는 대개 전체 표본 규모에서 -1을 취한 값을 자유도로 봄.
- 자유도가 증가함에 따라 t-분포는 표준정규분포와 거의 흡사해짐.
    - 자유도가 30을 넘을 때에는 거의 같다고 볼 수 있음.
    - 만약 자유도가 무한대라면 이는 표준정규분포라고 할 수 있음.
- 평균에 대한 신뢰구간은 오차한계를 비율에 대한 신뢰구간에서처럼 $z(se)$가 아니라 $t(se)$로 갖게 됨.

한 번 t-분포와 z-분포, 그리고 표본 규모에 따른 차이를 한 번 살펴보도록 하겠습니다.

```
x <- seq(-4, 4, length=1000)
hx <- dnorm(x)
degf <- c(1, 5, 30)
colors <- c("green", "blue", "red", "black")
labels <- c("n = 1", "n = 5", "n = 30", "normal")
plot(x, hx, type="l", lty=1,
     xlab="x-values",
     ylab="density", main="Comparison of t-distributions to Std.normal distribution")
abline(v = 0)
for (i in 1:3){
  lines(x, dt(x,degf[i]), lwd=2, col=colors[i])
  }
legend("topright", inset=.05, title="Distribution",
       labels, lwd=2, lty=c(1, 1, 1, 1), col=colors)
```

[2] t-분포를 이용한 방법은 더블린의 기네스 맥주공장에서 일하던 통계학자 William Gosset에 의해서 1908년에 고안되었습니다. 회사 정책 상 자신의 이름으로 논문을 내는 것이 불가능했기 때문에, Gosset은 가명으로 Student라는 이름을 이용하여 논문을 작성하였습니다. 때로 `Student's t`라고 불리는 이유는 그 때문입니다. 공장에서 Gosset에게는 테스트할 맥주 표본이 매우 소수만 제공되었기 때문에, 그는 그 표본사례를 가지고 기존의 표준오차 공식의 정규 z-score에 사용할 수 없다는 것을 알게 되었습니다. 그 결과가 소수의 사례로 구성된 꼬리가 정규표준분포에 비해 약간 더 두꺼운 t-분포입니다.



정규분포를 띄는 모집단으로부터 무작위 표본을 추출했다고 할 때, 95%의 확률로 모집단 평균 μ에 대한 신뢰구간은

$$\bar{y} \pm t_{.025}(se) \text{ , with } se = \frac{s}{\sqrt{n}}$$

이라고 할 수 있습니다. 이때 t-score에 따라 df $= n - 1$입니다.

모집단의 정규성을 가정하는 것은 표집분포가 표본규모에 상관없이 종의 형태를 띌 것이라는 것을 보장해주기 때문입니다. Figure 7.1을 보시면 이해가 더 쉬우리라 생각됩니다.

예제로 한 번 평균에 대한 신뢰구간을 살펴보도록 하겠습니다. 관측치가 1,467개라고 할 때, 친한 친구 수의 평균에 대한 95%의 신뢰구간이 $(6.8, 8.0)$이라고 합시다. 그렇다면, 다음 중 어떤 해석이 올바른 해석이라고 할 수 있을까요?

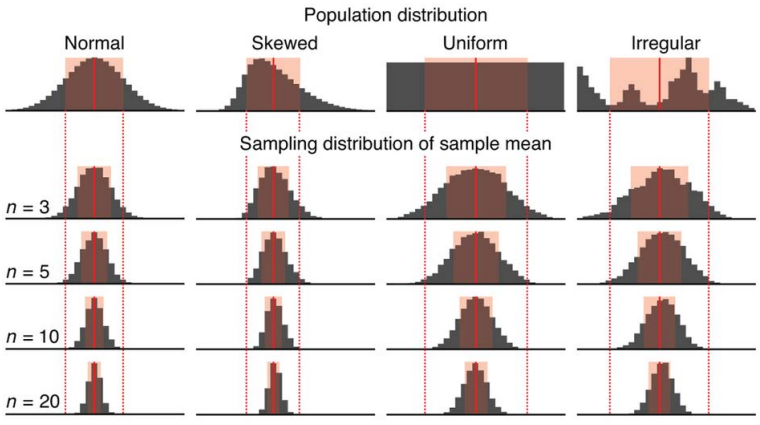


Figure 7.1: Population Distribution

1. 우리는 95%의 확률로 $\bar{y}$가 6.8과 9.0 사이에 존재한다고 할 수 있다.
2. y가 친한 친구의 수라고 (이 표본에서) 할 때, 그 값이 95%의 확률로 6.8에서 8.0 사이에 위치한다고 할 수 있다.
3. 만약 표본규모가 1,467인 무작위 표본들이 반복적으로 추출된다면, 95%의 확률로 $\bar{y}$가 6.8과 8.0 사이에 떨어질 것이라고 말할 수 있다.
4. 만약 표본규모가 1,467인 무작위 표본들이 반복적으로 추출된다면, 장기적으로 계산된 신뢰구간들 중 약 95%가 모집단 평균(μ)을 포함할 것이다.

정답은 4번입니다. 오해하시면 안 되는 게, 신뢰구간은 어디까지나 표집(sampling)의 개념에 바탕을 두고 있습니다. 하나의 표본에서의 평균이 어느 구간에 존재할 것이라고 주장하는 것이 아니라는 점을 유념하셔야 합니다.

μ에 대한 신뢰구간을 구하는 방법은 모집단의 정규분포 가정이 위배되는 상황에서도 매우 유용합니다. 우리는 직관적으로 신뢰수준이 높을수록 신뢰구간이 넓어진다는 것과 표본규모가 커질수록 더 좁은 신뢰구간을 가진다는 것을 알 수 있습니다.

표집을 수없이 반복해보면, 장기적으로 95% 신뢰수준에서 우리는 μ에 대한 신뢰구간이 95%의 확률로 실제 모집단의 평균, μ를 포함하고 있을 것이라고 주장할 수 있습니다. 다시 한 번 강조하지만, 개별 표본의 신뢰구간이 중요한 것이 아니라 이론적으로 모집단에서 수많은 표본들을 뽑아낼 수 있다고 할 때, 그 표본들로부터 얻어내는 통계치의 분포, 표집분포(sampling distribution)이 얼마나 모집단의 모수 추정에 도움이 되는가가 중요합니다. 모집단의 정규성 가정은 수많은 표집을 통해 완화될 수 있습니다.

7.2 유의성 검정(Significnace Tests)

추정, 신뢰수준, 신뢰구간의 논의에 이어서 이 절에서는 유의성 검정에 대한 이야기를 해보고자 합니다. 구체적으로는 유의성 검정을 구성하는 다섯 가지 요소들과 평균과 비율에 대한 유의성 검정, 그리고 통계적 유의성을 확인하는 데 주의해야할 오류의 유형들과 유의성 검정의 한계들을 다룰 것입니다.

7.2.1 유의성 검정을 이루는 다섯 가지 부분

유의성 검정의 다섯 요소들을 살펴보기에 앞서 이해해야 하는 것은 가설(hypothesis)입니다. 가설이란 연구의 변수들과 관련하여 모수로 표현된 모집단에 대한 예측을 말합니다. 즉, 모집단의 평균, 비율, 또는 상관관계가 어떠할 것이라는 기대를 보여줍니다. 유의성 검정은 데이터를 이용하여 가설로 예측된 값과 모수에 대한 표본 점추정값들을 비교함으로써 가설을 평가하는 절차를 말합니다.

이러한 유의성 검정을 이루는 다섯 가지 요소로는 첫 번째, 가정(assumptions)이 있습니다. 먼저 우리는 우리가 가진 데이터가 정량적(quantitative)인지 혹은 분류형(categorical)인지 알아야 합니다. 그리고 표집 방법으로는 무작위화(randomization)이 가정됩니다. 모집단의 분포는 정규분포(normal distribution)일 것으로 가정되며, 표본의 규모가 커질수록 검정의 타당성도 증가할 것으로 가정합니다.

두 번째 요소는 가설입니다. 우리는 영가설(null hypothesis, H_0)과 대안가설(alternative hypothesis, H_A)를 구분합니다. 영가설이란 모수가 정확하게 ㅇ떠한 값을 가질 것이라는 진술입니다. 대개 유의성 검정에서 영가설은 우리가 기대하는 설명변수의 효과가 없을 것(no effect)이라고 구성됩니다. 반면, 대안가설은 모수의 값이 어떠한 범주의 값에 속할 것이라는 진술로, "효과가 있을 것"이라는 기대를 보여줍니다.

세 번째는 검정통계치(test statistic)입니다. 우리는 가지고 있는 데이터와 영가설의 예측을 비교합니다. 주로 그 방법은 표본의 점추정치와 이를 둘러싼 표준오차의 구간—신뢰구간이 영가설에서 주장하는 모수값을 포함하고 있는지를 살펴보는 것입니다.

네 번째는 유의수준—P-값입니다. P-값은 영가설이 맞을 근거를 확률로 측정한 것입니다. 영가설이 맞을 확률은 곧 대안가설을 겸험적 결과가 지지하지 않을 확률과 같습니다. P-value가 작을수록 영가설이 맞을 확률, 경험적 근거가 적다는 것이므로 이는 영가설을 기각할 가능성이 커진다는 것을 의미합니다.

마지막으로는 결론(conclusion입니다. 만약 결정할 필요가 없이 명백하게 영가설을 지지하는 근거가 미비하다면, P-값을 보고하고 결과를 해석합니다. 하지만 만약 결정이 필요하다면, 우리는 대개 어떤 기준에서 통계적으로 유의하다고 할지 기준점(cutoff points)을 설정하고(예를 들어, 1%나 5% 처럼), P-값이 그 기준점보다 작을 경우에 영가설을 기각합니다. 가장 흔히 받아들여지는 기준점은 0.05입니다.

- 대개는 P-값이 0.05보다 작거나 같을 때, 0.05 수준에서 유의하다라고 검정 결과를 보고합니다.

만약 P-값이 충분히 작지 않다면, 우리는 영가설을 기각하는 것을 실패하게 됩니다. 그러면 우리는 영가설이 필연적으로 참이지는 않지만, 그렇다고 그것이 아니라고 할 근거도 충분하지 않으므로 대안가설에 대해 주장하기가 어려워집니다.

7.2.1.1 평균에 대한 유의성 검정

예시를 통해서 평균에 대한 유의성 검정을 살펴보겠습니다. 만약에 A라고 하는 다이어트 방법이 있고, 그 다이어트 방법을 기점으로 몸무게를 측정했다고 해보겠습니다. 그렇다면 y는 몸무개의 변화라고 할 수 있을 것입니다. 그리고 표본의 크기는 17이며, 데이터가 아래와 같이 있다고 가정하겠습니다.

```
y <- c(11.4, 11.0, 5.5, 9.4, 13.6, -2.9, -0.1, 7.4, 21.5,
       -5.3, -3.8, 13.4, 13.1, 9.0, 3.9, 5.7, 10.7)
n <- length(y)
mean <- mean(y)
sd <- sd(y)
```

과연 A라는 다이어트 방법이 효과가 있는지 그 근거는 어떻게 확인할 수 있을까요?

- 먼저, μ를 모집단의 평균 몸무게 변화라고 하겠습니다.
- 그리고 우리는 영가설, $H_0 : \mu = 0$ (효과가 없음)을 $H_A : \mu \neq 0$ (효과가 없지 않음)이라는 연구가설에 대해 검정해보겠습니다.

주어진 자료의 표본규모(n), 평균(mean), 그리고 표준편차(sd)를 이용해서 간단하게 신뢰구간을 계산해주겠습니다. 모집단 평균에 대한 신뢰구간과 표준오차는 같기 때문에,

$$se = s/\sqrt{n}$$

로 계산해주면,

```
se <- sd / sqrt(n)
se
```

```
## [1] 1.73593
```

위와 같은 결과를 얻게 되고, 주어진 자유도(17 -1 = 16)에서 검정통계량은 $t = \frac{\bar{y}-\mu_0}{se}$ 라고 할 수 있습니다. 따라서 그 결과는 아래와 같이 계산할 수 있습니다.

```
t <- (mean - 0)/se
t
```

```
## [1] 4.184908
```

그렇다면 이 t-값은 유의수준과 어떠한 관계에 있을까요? 우리는 t-값이 계산한 결과보다 클 확률을 양측꼬리 검정으로 계산해볼 수 있습니다.

```
2 * pt(-abs(t), df = n - 1)
```

```
## [1] 0.0007002531
```

해석하자면, 만약 영가설이 참이라면 영가설값(0)으로부터 최소 약 4.18배의 표준오차 범위 내에서 표본평균을 얻을 확률이 약 0.0007이라는 것이라고 할 수 있습니다. 결국, 우리는 이를 통해 모집단 평균이 0과 다를 것이라는 강력한 근거를 확보하게 되는 것입니다.

- 양측꼬리 검정에서 P-값이 0.05보다 작거나 같을 때, 모집단 평균에 대한 95% 신뢰구간은 모집단 평균에 대한 영가설의 값(예를 들어 효과가 없음을 보여주는 0)을 포함하지 않을 것입니다.
- 만약 P-값이 0.05보다 크다면, 95% 신뢰구간은 필연적으로 모집단 평균에 대한 영가설의 값을 포함하고 있을 것입니다.
- 이러한 점에서 신뢰구간은 실제 모집단 평균의 값에 대한 더 많은 정보를 제공합니다.

그렇다면 단측꼬리 검정은 어떠한 차이가 있을까요? 단측꼬리 검정은 모집단 평균에 대한 영가설에 방향성을 부과한 것이라고 이해할 수 있습니다. 예를 들어, 앞서의 예제에서 다이어트 방법 A에 따라 몸무게의 변화의 평균은 0보다 클 것($H_0 : \mu > 0$)

이라는 것입니다. 만약 t-값이 우측꼬리에서 멀어질수록 이 가설을 데이터가 지지하므로, P-값은 우측꼬리에 대한 확률이라고 할 수 있습니다.

- $n = 4$(df $= 3$)일 때, $t = 2.0$라고 해보겠습니다. 그렇다면 P-값은 $P = P(t > 2.0)$이므로 0.07이라고 할 수 있습니다.
- 그렇다면 대안가설($H_A : \mu < 0$)의 확률은 좌측꼬리를 기준으로 P-값이 $P = P(t < 2.0)$이므로 0.93이 됩니다.

일반적으로 양측꼬리 검정이 더 흔하게 사용됩니다. 왜냐하면 관계 양상의 존재—인과적 효과를 살펴보기 위해 가설을 수립하는 경우가 더 많기 때문입니다.

이렇게 영가설과 대안가설을 수립하고 나서는 "결정"을 해야합니다. 유의수준(significance level)이라 불리는 α-level은 고정된 수입니다. 우리는 이 유의수준에 근거하여,

- P-값이 α 값보다 작거나 같으면 "영가설을 기각한다." 또는,
- P-값이 α 값보다 크면 "영가설을 기각하지 않는다."라는 결정을 내릴 수 있습니다.

우리는 "영가설을 채택한다"라는 표현보다는 "**영가설을 기각하지 않는다**"라고 말합니다. 왜냐하면 영가설이 맞다는 것이 아니라 단지 우리가 확인한 것은 영가설의 값이 신뢰구간에 포함될 확률이 기대보다 높게 나타났을 따름입니다. 우리는 과연 어떤 설명변수의 효과가 존재하는지, 존재하지 않는지에 대해서 확언할 수가 없습니다.

한편, 이와 같은 유의성 검정에 대한 표본 규모의 효과를 살펴볼 필요가 있습니다. 대개는 관측치가 30개를 초과할 때, CLT로 인하여 정규분포에 대한 가정은 크게 중요하지 않다고 봅니다. 다만, 표본 규모가 30개보다 작은, 소규모 표본일때는 양측꼬리 t-검정을 수행하는 것이 정규분포 가정이 위배되는 것을 일부 방어할 수 있습니다. t-분포는 소규모 사례 분포에 대한 분포를 가정하고 있으니까요.

표본의 규모가 커질수록 검정 통계치도 더 커집니다. 왜냐하면 공식 상 분모가 표준오차인데, 표준오차는 표본의 규모가 커질수록 감소하기 때문입니다. 따라서 표본의 규모가 커질수록 우리는 자동적으로 P-값이 작아지는 것을 확인할 수 있습니다. 데이터가 많을수록 근거를 확보하기가 쉬워진다는 것을 의미합니다. **그러나 표본 규모가 클 때, 통계적 유의성은 결코 실질적 유의성과 같은 의미는 아닙니다.** 통계적으로 유의하다고 해서 과연 그 효과가 실질적으로 유의미한 것일까요? 예시를 들어보겠습니다.

아까와 같이 몸무게 변화에 대해 평균 $1, 0$, 표준편차 2.0, 사례 수가 400개라고 해보겠습니다. 그렇다면 표준오차는 $2.0/\sqrt{400} = 0.1$가 되고, t-값은 $(1.0 - 0)/0.1 = 10.0$이 됩니다. 그럼 P-값은 $0.00000\cdots$으로 엄청 작게 나타날 것입니다. 그리고 95% 신뢰구간은 $(0.8, 1.2)$입니다. 결과적으로는 다이어트 방법 `A`는 몸무게 변화에 긍정적인 효과가 있는 것으로 나타났지만 실질적으로 그 효과는 매우 작습니다. 따라서 통계적으로는 유의미할지라도 실질적으로는 유의하지 않을 수도 있습니다.

7.2.1.2 비율에 대한 유의성 검정

비율에 대한 유의성 검정도 평균과 마찬가지로 다섯 가지 요소들을 중심으로 접근해볼 수 있습니다. 먼저 비율에 대한 유의성 검정의 가정은 데이터가 명목형-분류형 변수일 것이며, 마찬가지로 무작위로 표집된 데이터여야 하고, 표본의 규모가 어느 정도 보장되어야 합니다(large-sample).

둘째로, 영가설에 있어서는 모비율($H_0 : \pi = \pi_0$)에 대한 기대를 설정하고, 대안가설로는 양측검정일 경우에는 $H_A : \pi \neq \pi_0$을, 단측검정일 경우에는 $H_A : \pi > \pi_0, H_A : \pi < \pi_0$로 설정해줄 수 있습니다.

그 다음으로 검정통계치는 평균과는 다르게 z를 사용하며, $z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$로 구할 수 있습니다. 이때, $\sigma_{\hat{\pi}}$는 영가설에 대한 표준오차(se_0)로 $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$으로 계산이 가능합니다. 주의해야할 점은 앞서 신뢰구간의$se = \sqrt{\hat{\pi}(1 - \hat{\pi})}$과 다르다는 점입니다. 검정통계치의 논리는 (모수에 대한 추정값 - 영가설 값)/(표준오차)로, 효과가 있다면 표집으로 인해 우연히 만들어질 변동(variation)인 표준오차보다 명백히 큰 값이 분자에 자리하게 되어 큰 z값을 생산할 것이라는 점입니다.

P-값은 세 가지 경우의 수를 생각해볼 수 있습니다.

- $H_A : \pi \neq \pi_0, P =$ 표준정규분포에서 양측꼬리 확률을 의미
- $H_A : \pi > \pi_0, P =$ 표준정규분포에서 우측꼬리 확률을 의미
- $H_A : \pi < \pi_0, P =$ 표준정규분포에서 좌측꼬리 확률을 의미

따라서 결론은 평균에 대한 검정과 다르지 않습니다. 예를 들어, 만약 P-값이 유의수준(α-값)보다 작거나 같다면, 영가설을 기각할 근거를 가지게 되는 것입니다.

7.2.2 오류의 유형 (Types of Errors)

검정에 있어서 우리는 유의수준과 기각역(rejection region)을 결정해야 합니다.

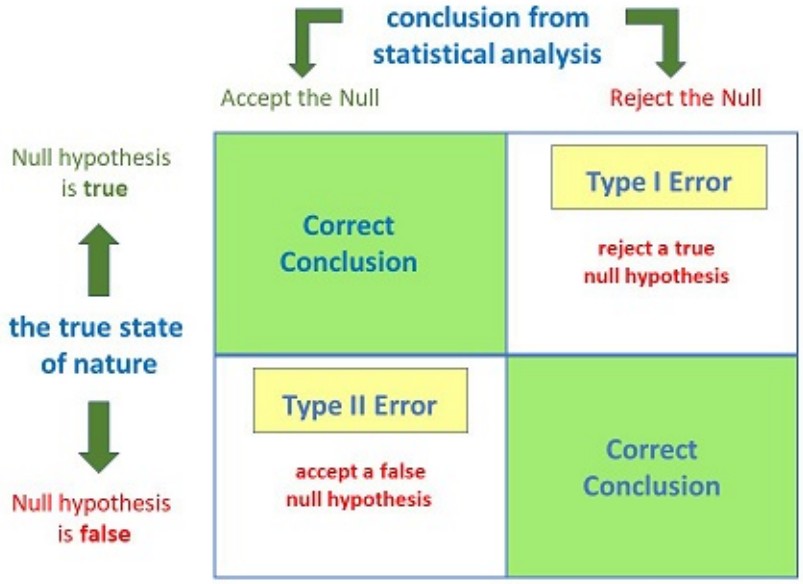
- 유의수준, α-값이란 P-값이 "이 기준"에 달하지 못할 때, 우리는 영가설을 기각한다고 미리 특정해둔 일종의 "허들"입니다. 전형적으로 0.05, 혹은 0.01 선에서 결정하고는 합니다.

Table 7.1: 검정결과의 결정

P-값	영가설 결론	대안가설 결론
$p \leq .05$	기각	채택
$p > .05$	기각하지 않음	채택하지 않음

- 기각역이란 우리가 영가설을 기각하기 위한 검정통계치의 값을 의미합니다. 예를 들어, 양측꼬리 검정일 경우 $\alpha = 0.05$라고 한다면 우리는 H_0 if $|z| \geq 1.96$일 경우에 영가설을 기각합니다.

문제는 이러한 우리의 결정이 불확실성에 기초하고 있기 때문에 오류가 존재할 수 있다는 것입니다. 크게 두 종류의 오류를 범할 수 있습니다.



- 제1종오류(Type I Error; false positive)
  - 영가설이 사실임에도 불구하고 이를 기각하는 오류
  - 간단하게 말하면 임신하지 않았는데 임신하였다고 잘못 진단하는 경우와도 같습니다.
  - 제1종오류를 범할 확률은 유의수준과 같습니다.
  - 전통적으로 검정에 있어서 효과가 없다고 의심하는 것을 더 중요시하기 때문에 작은 유의수준값을 견지하나, 유의수준이 너무 작을 경우에는 제2종오류를 범할 확률, β가 높아지게 됩니다.
- 제2종오류(Type II Error; false negative)
  - 영가설이 거짓임에도 불구하고 이를 기각하지 못하는 오류
  - 임신한 사람에게 임신하지 않았다고 진단하는 경우와 같습니다.
  - 제2종오류를 범할 확률, β는 모수의 참값에 좌우됩니다. 모수의 참값이 영가설의 값으로부터 멀어질수록, 영가설을 기각하기는 쉬워지고 β도 감소하게 됩니다.
  - 실질적으로 표본 규모가 충분히 크다면, 제2종오류는 감소합니다.

7.3 선형회귀와 상관관계

제6장에서 간단하게 언급되었던 선형관계, 확률모델, 선형회귀모델, 선형 관계를 측정하기 위한 상관관계, 마지막으로 기울기와 상관관계에 대한 추론 등의 이야기를 이어나가도록 하겠습니다.

7.3.1 선형관계

선형관계를 살펴보기 위해서 우리는 정량적인 데이터를 가져야만 합니다. 우리는 정량적 변수들을 가지고 과연 이 둘 간에 관계가 있는지(독립성 검정), 있다면 그 관계의 강도는 어떻게 되는지(상관관계), 나아가 x를 이용해 y를 예측하는 방식으로 관계의 특성을 어떻게 서술할 수 있는지(회귀식, 잔차)를 살펴볼 것입니다.

먼저, 선형관계를 우리는 선형함수(linear function)로 나타낼 수 있습니다. 간단히 말하면 두 변수 간의 관계를 반듯한 직선의 관계로 보여줄 수 있다는 것을 의미합니다. 수학적으로는 $y = \alpha + \beta x$라고 할 수 있겠네요. 이 식은 우리가 y를 x에 관한 기울기 β와 절편 α의 선형함수로 표현한 것입니다. 이 함수에 따르면 x의 한 단위 증가는 y를 β 단위만큼 증가시킵니다.

- $\beta > 0$ 선의 기울기가 우상향하는 정의 관계(positive relationship)
- $\beta = 0$ 선의 기울기가 수평(y는 x와 관계가 없음)
- $\beta < 0$ 선을 기울기가 우하향하는 부의 관계(negative relation)

경제 수준과 이산화탄소(CO_2) 배출 변수를 예제로 선형 관계에 대해 좀 더 구체적으로 살펴보겠습니다.

- 종속변수(y)는 이산화탄소 배출량입니다.
- 설명변수(x)는 1인당 GDP입니다.
- 이 둘의 선형관계를 대략적으로 $y = 0.42 + 0.31x$로 나타낼 수 있다고 해보겠습니다.
  - $x = 0$일 때, 주어진 선형회귀모델로 예측할 수 있는 이산화탄소 배출량은 $y = 0.42 + 0.31(0) = 0.42$이라고 할 수 있습니다.
    - 물론, 현실적이지는 않습니다. 1인당 GDP가 0인 경우는 없기 때문입니다.
  - 한편, $x = 39.7$일때, 예측된 이산화탄소 배출량은 $y = 0.42 + 0.31(39.7) = 12.7$가 됩니다.
    - 그런데, 실제 배출량은 19.8이라고 해봅시다.
    - 1인당 GDP가 1천 달러씩 증가할 때마다, 예측되는 이산화탄소 배출량은 0.31 메가톤씩 증가할 것이라고 예측됩니다.
    - 그러나 이러한 선형 회귀식은 어디까지나 개략적이라고 할 수 있습니다. x와 y간의 상관관계는 1이 아니라 0.64에 불과하기 때문입니다.

왜 이런 결과가 나왔을까요? 먼저, 우리는 변수 코딩에 따른 효과를 생각해볼 수 있습니다. 기울기와 절편은 측정단위에 따라서 달라질 수 있습니다. 만약 설명변수인 1인당 GDP를 1천 달러가 아니라 달러 단위로 측정했다면 어떻게 될까요? 아마 회귀식은 $y = 0.42 + 0.00031x$으로 추정될 것입니다.

- 이 경우 달러로 측정된 x의 한 단위 변화는 1천 달러로 측정된 x의 한 단위 변화의 효과의 1/1000밖에 되지 않습니다. 따라서 기울기는 0.001배가 됩니다. 만약 y인 이산화탄소 배출량이 메가톤(=1천 킬로그램)이 아니라 그냥 킬로그램으로 측정되었다면 어떻게 될까요? 회귀식은 $y = 1000(0.42 + 0.00031x) = 420 + 0.31x$이 될 것입니다.
- 말 그대로 단위가 달러에서 파운드로 바뀐다고 생각해도 이러한 효과의 변화가 나타나는 것은 마찬가지입니다.

7.3.2 확률모델(Probability model)

실질적으로 y와 x의 관계는 완벽하지 않습니다. 왜냐하면 y가 x에 의해서 완벽하게 설명될 수 없기 때문입니다. x로는 설명할 수 없는 y의 속성이 존재하고, 현실에서 우리는 그 외부의 요인들을 모두 관측하여 계량 및 측정할 수는 없습니다.

- 우리는 $\alpha + \beta x$가 x의 함수적 작용으로서 나타나는 y 값의 평균을 보여준다고 봅니다.
- 평균은 사실 모집단에서 나타날 종속변수의 기대값($E(y)$)을 표본 수준에서 보여주는 것입니다.
  - y의 기대값($E(y)$)는 y의 확률분포의 평균으로 생각할 수 있습니다.
- 만약 y가 소득, x가 교육 연수라고 생각해보면, $E(y) = \alpha + \beta(12)$는 모집단에서 교육 연수가 12년차인 모든 사람들의 평균 소득을 보여줄 것이라고 기대됩니다.

정리하자면 회귀함수는 종속변수 y의 평균이 설명변수 x의 값에 따라서 어떻게 변화하는가를 보여주는 수학적 함수라고 할 수 있습니다. 그렇다면 왜 굳이 "회귀"(regression)라는 표현을 사용하는 것일까요?

선형회귀함수는 관계를 보여주는, 현실에 대한 단순화한 표현인 모델의 일부에 불과합니다. 만약 진정한 현실에서의 관계가 대략적으로 선형이라면 선형 모델로 그 관계를 보여주는 것은 문제가 없겠지만 현실세계의 모습이 매우 비선형적 관계를 보일 때에는 선형 모델을 이용할 때, 그 결과가 현실을 추론하는 데 도움을 줄 것이라고 기대하기 어려울 것입니다.

그렇다면 우리는 어떻게 선형식(linear equation)을 추정할까요? 제6장에서처럼 예비적 분석으로 산포도를 살펴보고, 선형 모델이 가능한지 여부를 확인하는 것이 하나의 방법입니다.

그 다음의 문제가 바로 어떻게 데이터에 가장 잘 들어맞는("best fit") 추세선을 선택할 수 있을까 입니다. 가장 흔히 사용되는 기준은 관측된 데이터로부터 추세선까지의 수직 거리의 제곱합이 최소화하는 선을 그린다는 것입니다. 우리는 이를 최소자승법(least squares)을 이용한 예측식(prediction equation)이라고 합니다.

수리적으로 이 선을 그리는 식은 다음과 같이 나타낼 수 있습니다.

- 모집단 수준에서의 절편 α의 추정값을 a로, 모집단 수준에서의 기울기 β의 추정값을 b로, 그리고 y의 기대값($E(y)$)에 대한 예측추정값을 표본에서 $\hat{y}$라고 보는 것이죠.
- 따라서 $\hat{y} = a + bx$라고 할 수 있고,
- 이때, b는 $b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$가 됩니다.
- 그리고 a는 $a = \bar{y} - b\bar{x}$로 나타낼 수 있습니다.

수리적으로 공식은 이렇게 구축할 수 있다고 하고, 과연 이 식의 의미는 무엇일까요? 무엇이 저렇게 식을 유도하게끔 한 걸까요? 개념적으로 한 번 식의 의미를 풀어보겠습니다.

- 만약 x와 y 값 모두가 평균보다 크거나 평균보다 작다면, $(x_i - \bar{x})(y_i - \bar{y})$는 양수가 됩니다.
- 앞서의 $a = \bar{y} - b\bar{x}$는 $\bar{y} = a + b\bar{x}$와 같습니다.
- 즉, x의 평균에서 y의 예측값은 y의 평균이 됩니다. 선형회귀선은 $(\bar{x}, \bar{y})$를 지나는 선이라고 할 수 있습니다.

관측치를 25개 가진 표본으로 산출한 $y = 18.41 + 1.67x$라는 식이 있다고 해보겠습니다. 해석해보면,

- x의 한 단위의 증가에 따라 y는 1.67만큼 증가하는 양상을 보입니다.
- y의 절편은 18.4로 x가 0일 때의 y의 값입니다.
- 기울기 $b = 1.67 > 0$이므로, 표본에 있어 변수들 간의 양의 관계를 보여줍니다.
- 그러나 표본의 규모가 작을 경우, 추세를 파악하기에는 변동량(variability)이 크기 때문에, 표본이 속할 것이라고 기대되는 모집단에서의 변수들의 관계 역시도 양의 관계라고 하기에는 어렵습니다.

예측오차(prediction errors), 잔차(residuals)에 대해서도 살펴보겠습니다. 하나의 관측치에 대해서 우리는 관측된 개별 y값, y_i와 주어진 x 값에서의 예측된 y값, $\hat{y}$ 간의 차이를 발견할 수 있습니다: $y - \hat{y}$. 이를 잔차라고 합니다. 회귀선을 그린 산포도를 예시로 생각해보면, 회귀선과 산포도의 각 관측치(점)들의 수직 거리를 의미합니다.

$x = 9, y = 37$인 한 점이 있다고 해보겠습니다. 위의 회귀식에서 $x = 9$일 때, $\hat{y} = 18.41 + 1.67(9) = 33.4$입니다. 따라서 이때의 잔차는 $37 - 33.4 = 3.6$이 될 것입니다.

우리는 잔차의 총합(과 평균)이 0일 것이라고 가정합니다. 이는 이론적으로 모집단에 대한 기대에서 도출된 것이라고 볼 수 있습니다. 모집단에서 우리는 y를 x를 가지고 설명할 수 있을 것이라고 생각합니다. King, Keohane, and Verba (1994)의 용어에 따르자면, x라는 체계적 요인(systematic factor)로 설명하고 남은 y는 y의 예측불가능한 비체계적 요인(nonsystematic errors)입니다. 비체계적이라는 것은 특별한 경향성을 가지지 않는다는 것으로 무작위(randomized)라는 것과 같고, 확률적이므로 그 확률분포의 평균은 0에 수렴하게 될 것이라는 기대입닏. $E(\epsilon) = 0$이라는 가정적 기대는 표본 수준에서 잔차의 총합 또는 평균이 0일 것이라는, 비체계적 요인들을 서로 그 효과가 상쇄(trade-offs)되어 종속변수에 유의미한 체계적 변화를 일으키지 않을 것이라는 주장의 근거가 됩니다.

앞서도 언급했다시피, 예측식은 최소자승법으로 그려집니다. 최소자승법이란 잔차의 제곱합이

$$SSE = \sum(y_i - \hat{y}_i)^2 = \sum[y_i - (a + bx_i)]^2$$

최소가 되도록 선을 그리는 방법입니다.

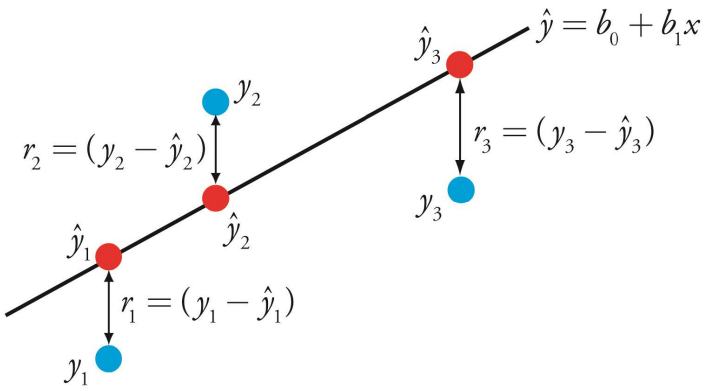


Figure 7.2: Residuals

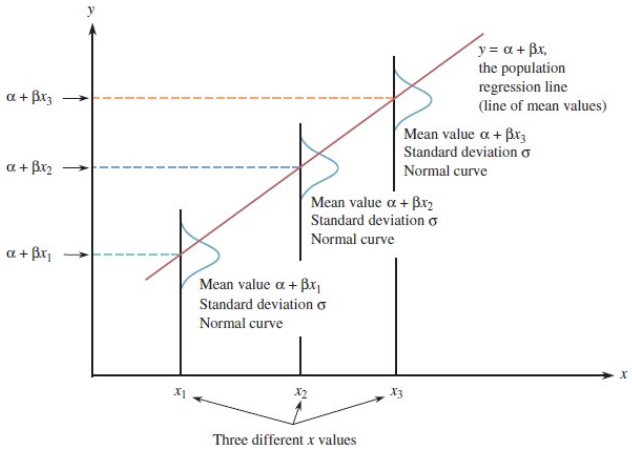


Figure 7.3: How to draw a regression line

7.3.3 선형회귀모델

선형회귀모델이 모집단 수준에서 $E(y) = \alpha + \beta x$라는 것을 다시 떠올려보겠습니다. 이 식은 각 x의 고정된 값에서 y의 조건분포의 평균들을 이으면 직선의 형태로 나타날 것이라는 의미입니다. 이러한 선형회귀모델은 또 다른 모수인 σ를 갖는데, σ는 조건분포의 변동성(variability)을 의미합니다. 즉, x값이 같다고 할 때, y값이 가질 수 있는 모든 가능성—변수이기 때문에 가질 수 있는 변동성을 의미합니다. y에 대한 조건적 표준편차의 추정값은 다음과 같이 구할 수 있습니다.

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-2)}}$$

```
knitr::include_graphics("./Chapters_pdfR/plot/regression_model2.jpg", dpi = 150)
```

선형회귀분석의 가정들은 흔히 가우스-마르코프(Gaus-Markov) 가정이라고 불리는데, 이 가정들에 대한 이론적 논의는 `Lv.2.Statistics` 자료들에서 다루도록 하겠습니다.

7.3.4 관계의 측정: 상관관계

마지막으로 살펴볼 내용은 바로 관계의 강도를 어떻게 측정하는가와 관련된 것입니다. 회귀식에서의 기울기는 x와 y 간의 관계의 방향을 보여줍니다. 하지만 기울기의 크기는 변수의 측정단위에 따라 달라질 수 있습니다. 따라서 우리는 관계의 강도를 살펴보기

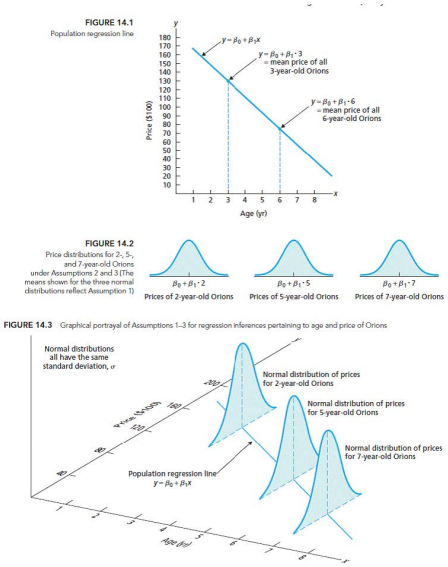


Figure 7.4: Assumptions and Models of Linear Regression

위해서 단위를 신경쓰지 않아도 되는 표준화된 기울기(standardized slope)인 상관관계—상관계수(correlation)을 볼 수 있습니다.

상관계수 r은 예측식의 기울기(b)의 맥락에서 $r = b(s_x/s_y)$로 나타낼 수 있습니다. s_x와 s_y는 x와 y의 한계 표본표준편차(marginal sample standard deviations)라고 할 수 있습니다.

7.3.4.1 상관계수의 특성

- 상관계수 r은 $s_x = s_y$일 때의 기울기 b인 표준화된 기울기입니다.
- 상관계수 r은 기울기 b와 같은 부호를 가지며, 표준화되었기 때문에 그 값은 $-1 \leq r \leq +1$ 사이에 위치합니다.
- 표본의 모든 관측치가 예측선 위에 정확하게 놓여있을 때, $r = 1$ 또는 $r = -1$로 나타나고 r은 선형관계의 강도를 보여줍니다.
- $b = 0$일 때, $r = 0$이며, 관계가 존재는 하지만 그것이 선형이 아닐 때 이런 양상이 나타날 수도 있습니다.
- r의 절대값이 커질수록 관계의 강도가 더 강하다는 것을 의미합니다.

```
knitr::include_graphics("./Chapters_pdfR/plot/correlation.jpg", dpi = 150)
```

상관계수의 절대값이 1보다 작다는 것은 예측이 평균으로 수렴한다는 것을 의미합니다.

- x가 한 단위 증가할 때, 예측된 y는 b만큼 변화한다는 의미입니다.
- x가 s_x 만큼 증가할 때, 예측된 y는 $s_x b = r s_y$ 만큼 변화한다는 의미입니다.
    - x가 1 표준편차만큼 증가할 때, y가 $r \times$ 표준편차만큼 변화할 것이라고 예측할 수 있습니다.

r^2는 예측오차가 비율적으로 얼마나 감소했는지를 보여주는 지표입니다. 간단하게 말하면 x가 얼마나 y를 잘 예측하는지를 보여주는 것입니다.

- 우리는 x를 이용하여 y를 예측하기 위해 회귀식을 구축합니다. 이때, 예측오차는 간단하게 잔차의 제곱합으로 보여줄 수 있습니다.: sum of squared errors, SSE $= \sum(y - \hat{y})^2$
- y를 x 없이 예측한다고 할 때, 가장 좋은 예측변수는 바로 y의 표본평균일 것입니다. 따라서 $bary$로 예측할 때의 예측오차에 대한 측정지표는 다음과 같습니다: total of squared errors, TSS $= \sum(y - \bar{y})^2$
- 즉, r^2는 x를 이용할 경우, $\bar{y}$로 예측할 때보다 SSE가 TSS에 비해 얼마나 감소하였는가를 보여주는 것입니다:

$$r^2 = \frac{TSS - SSE}{TSS} = \frac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

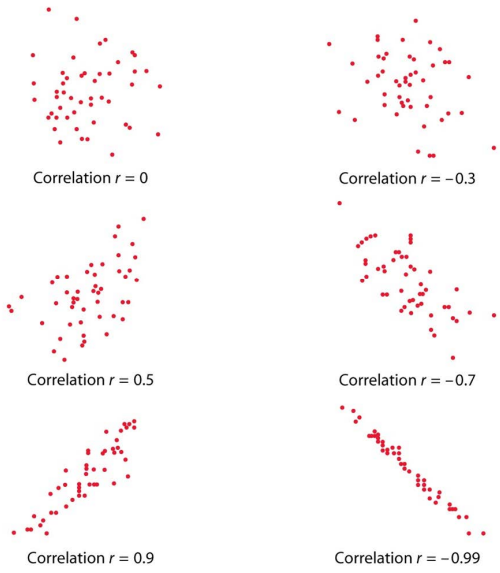


Figure 7.5: Assumptions and Models of Linear Regression

- 그리고 단순선형회귀모델에서 오차의 비율적 감소, r^2는 말 그대로 상관계수의 제곱과도 같습니다. r^2는 "알 스퀘어(r-squared)", 혹은 결정계수(coefficient of determination)이라고도 불립니다.
  - 당연히 r의 제곱이므로 값은 $0 \leq r^2 \leq 1$ 사이에 위치합니다.
  - 가능한 SSE의 최소값은 0입니다. 이 경우는 $r^2 = 1$인 경우로 모든 표본의 관측치들이 예측선 위에 올라와있는 것을 의미합니다.
  - 만약 $b = 0$이라고 한다면, $a = \bar{y} - b\bar{x} = \bar{y}$이고, 따라서 TSS = SSE가 되어 $r^2 = 0$가 됩니다.
  - r^2는 x, y의 측정단위에 영향을 받지 않습니다.

7.3.4.2 기울기와 상관계수에 대한 추론

통계적 추론에 대한 가정을 생각해보겠습니다. 먼저, 우리에게는 무작위로 표집된 표본이 필요합니다. 그리고 y의 평균은 x와 선형으로 연계되야 합니다: $y = \alpha + \beta x$. 그리고 조건적 표준편차(s), σ는 각 x 값에서 동일해야하고, 각 x 값에서의 y의 조건분포는 정규성을 띄어야 합니다.

하지만 실질적으로 이러한 가정들 중 어느것도 완벽하게 만족되기 힘듭니다. 다만 우리는 그 중 무작위성과 선형성 가정만큼은 매우 중요한 것으로 간주합니다.

그렇다면 x와 y가 독립적인지 여부를 어떻게 검정할까요? 우리가 기대하는 모수란 회귀모델에서 모집단의 기울기를 보여주는 β일 것입니다. 그고 그 모수에 대한 추정치는 최소자승법에 따라 추정된 기울기, b입니다. 또한 우리는 표준오차도 추정할 수 있습니다:

$$se = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} - \frac{s}{s_x \sqrt{(n-1)}}$$

. 그리고 공식에서 확인할 수 있듯이 표본의 규모(n)이 증가할수록 표준오차는 대개 감소합니다.

우리는 x와 y가 서로 독립적일 것, 즉 관계가 없을 것이라는 영가설($H_0 : \beta = 0$)을 세웁니다. 그리고 대안가설은 양측검정—방향성을 고려하지 않을 때에는 $H_A : \beta \neq 0$으로 수립하며, 만약 방향성을 고려하는 단측검정일 경우에는 $H_A : \beta > 0$, 또는 $H_A : \beta < 0$으로 수립합니다.

이러한 영가설에 대한 검정통계치, t는 $t = (b - 0)/se$이며, 이때의 자유도는 절편과 기울기 변수를 제외한 값(df = $n - 2$)입니다.

통계적 추론에서 우리는 기울기 β에 대한 신뢰구간을 다시 한 번 생각해볼 수 있습니다.

- β에 대한 신뢰구간은 $b \pm t(se)$의 형태로 나타낼 수 있습니다.
- 예를 들어, $b = 1.666, se = 0.692$, 그리고 df= 8이라고 해보겠습니다.

- β에 대한 95% 신뢰구간에서 t-score는 $t_{0.025} = 2.306$입니.
- 따라서 신뢰구간은 $1.666 \pm 2.306(0.692)$으로 $(0.07, 3.26)$이 됩니다.
- 이때 우리는 주어진 신뢰구간에서 변수들의 모집단에서의 관계가 양의 관계라고 결론을 내릴 수 있습니다.
  * 양측검정에서 P-값은 0.04로 0.05 신뢰수준에서 영가설($H_0 : \beta = 0$)을 기각하므로 우리는 관계가 존재할 것이라고 결론내릴 수 있습니다.
  * 이처럼 β에 대한 95% 시뢰구간은 0을 포함하지 않을 때, 우리는 β가 0이 아닐 것—설명변수 x의 종속변수 y에 대한 효과가 없는 것이 아닐 것(reject the null)이라고 말할 수 있는 것입니다.

7.3.5 정리하며

선형회귀모델에서 어떤 가정들이 중요할까요?

1. 선형회귀는 모델입니다.

- 따라서 우리는 변수들 간의 관계가 완벽하게 선형일 것이라고 기대할 수는 없습니다.
- 그러나 그러한 선형 관계는 실제에 대한 단순화된 개략(approximation)을 보여줄 수는 있습니다.
- 만약 실제 변수들의 관계가 U자형이라면, 특정한 관계는 존재할지 모르지만 선형회귀모델로는 P-값이 작게 나타나지 않을 것입니다.
- 따라서 이 근본적 가정을 확인하기 위햇 우리는 항상 산포도와 같은 변수들의 관계 양상을 직관적으로 살펴볼 필요가 있습니다.

2. 각 x값에 대한 표준편차 σ가 동일할 때, y의 조건분포는 정규적입니다.

- 현실세계에서 이 가정은 완벽하게 들어맞기 힘듭니다.
- 표본의 규모가 충분히 클 때, 정규성 가정은 CLT에 따라서 이론적으로 만족될 수 있기 때문에 그다지 중요하지는 않습니다.
- 만약 σ가 동일하다는 가정이 위배된다면, 최소자승법보다 더 작은 표준오차를 가진 추정값들이 더 효율적일 수 있습니다.
- 그러나 통상적인 추론방법은 이러한 가정들이 완벽하게 충족되지 않더라도 "대략적으로" 타당하다고 할 수 있습니다.

3. x값의 관측된 범주 밖의 외삽(extrapolation)의 문제는 위험합니다.

- y가 고등학교 성적이고 x가 주간 TV 시청시간이라고 해보겠습니다.
- 이때, 우리가 가진 회귀식은 $\hat{y} = 3.44 - 0.33x$라고 해봅시다.
- 하지만 $x = 100$일때, 고등학교 성적 0.44를 공식 상으로 예측할 수는 있지만 이는 말이 되지 않습니다.
- 따라서 외삽의 문제에 따른 성급한 추론의 일반화를 경계해야할 필요가 있습니다.

4. 마찬가지로 통계적 유의성과 실질적 유의성에 대해 고민해야 합니다.