

데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

뭣이 중헌디? - 인과관계 알아보기

1. 인과관계란?

대학에 가면 뇌종양에 걸릴 확률이 높아진다?

- 2016년 400만명의 스웨덴 남녀를 대상으로 한 연구에서 납 세기록과 건강관련 기록을 분석한 결과 사회경제적 지위가 높은 남자일수록 뇌종양에 걸릴 경우가 더 많았다.
- 이 연구결과는 한 언론에 의해서 “왜 대학에 가면 뇌종양에 걸릴 위험이 커지는가?”라는 기사로 소개되었다.
- 부자일수록 진단검사를 더 많이 받을 수 있기때문에 생기는 확인편향(ascertainment bias)일 수 있다.

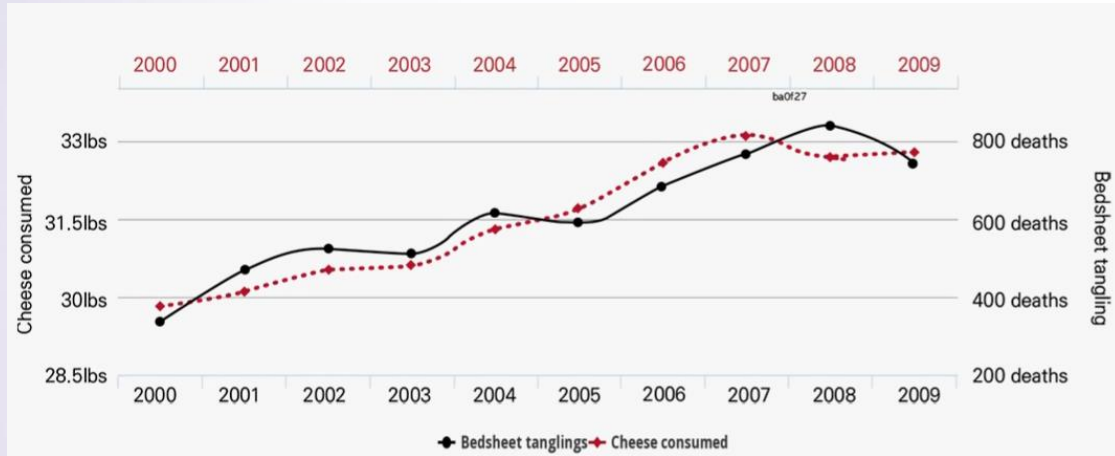
대학에 가면 뇌종양에 걸릴 확률이 높아진다?

- 비슷한 사례로 2013년 국내 모 언론사가 국민건강보험공단의 자료를 이용한 분석결과를 “중증질환, 부자가 더 잘 걸린다”라는 제목의 기사로 소개했다.
- 정확한 의미는 “부자들이 중증질환 병원 이용을 더 자주한다”. 즉 저소득층의 경우 의료비 부담때문에 중증질환이라도 병원에 가지 못하는 경우가 있다.

상관관계가 인과관계를 의미하지 않는다

- 피어슨 상관계수의 창안자인 피어슨은 본인의 상관계수가 인과관계를 의미하지 않는다는 점을 명확히 했다.
- “Spurious correlations”이라는 웹사이트 (<https://www.tylervigen.com/spurious-correlations>)에서는 특이한 상관관계를 보여주는 다양한 예제를 제시한다.

상관관계가 인과관계를 의미하지 않는다



[1인당 치즈 소비량과 침대 시트에 얽혀 사망한 사람들의 수]
(Spurious correlations)

- 위의 그림은 1인당 치즈 소비량과 침대 시트에 얽혀 사망한 사람들의 숫자를 보여주고 있다. 이 둘 사이의 피어슨 상관관계수 값은 0.95이다!

선풍기를 틀고 자면 죽는다?

- 한국에만 있는 괴담 중 하나는 “선풍기를 틀고 자면 죽는다”가 있다. (그 외에도 혈액형 성격 설이 있지만 일본에서도 믿고 있다고 알려져 있어서 우리나라에만 있는 괴담은 아니다.)
- 만약 여름에 사망한 사람 중 선풍기를 틀고 자고 있었던 사람을 상당수 찾을 수 있다. 하지만 그 사실이 인과관계를 증명하는 것은 아니다! 사망하기 전에 저녁을 먹고 잔 사람들이 대다수라고 저녁밥이 사망원인이 될 수 없는 것과 같은 이치이다.
- 이러한 가설을 확인하기 위해서는 어떻게 해야 할까?
 - 임상시험을 통하여 확인해야 한다.
 - 결과가 반복적으로 여러 연구에서 확인되어야 한다.

스타틴은 심장마비와 뇌졸중을 예방하는데 도움을 주는가?

- 스타틴을 콜레스테롤을 낮추어서 심장마비와 뇌졸중의 위험을 줄이는 약이다.
- 스타틴의 이러한 효과에 대한 검증은 대규모 임상시험을 통해서 이루어진다.
- 임상시험은 다음과 같은 원칙이 필요하다.
 1. 대조군이 반드시 있어야 한다.
 2. 실험군과 대조군의 배정은 임의로 이루어져야 한다. 즉 Randomized Controlled Trial(RCT)이어야 한다.
 3. 분석은 맨 처음 할당된 그룹별로 실시되어야 한다.(intention to treat 원칙)

스타틴은 심장마비와 뇌졸중을 예방하는데 도움을 주는가?

4. 참가자들은 가능하면 본인이 어떤 그룹에 속하는지 몰라야 한다.
5. 각 그룹은 동일하게 다루어져야 한다.
6. 최종 결과의 평가자 (의사)역시 연구대상이 어떤 그룹에 속하는지 몰라야 한다.
7. 실험에 참가한 모든 사람은 (가능한 한) 끝까지 추적해야 한다.
8. 한 연구에만 의존하면 안 된다.
9. 증거를 메타분석(meta analysis)를 통해서 체계적으로 검토해야 한다.

스타틴은 심장마비와 뇌졸중을 예방하는데 도움을 주는가?

- 1990년대 후반에 실시한 영국의 심장보호연구 (Heart Protection Study, HPS)는 심장마비나 뇌졸중 위험이 높은 2만 536명의 사람들에게 40밀리그램의 스타틴과 위약 중 하나를 임의로 할당해 매일 먹게 했다. HPS 실시 5년 뒤의 결과는 다음과 같다.

사망원인	위약그룹	스타틴그룹	스타틴 배정 시 상대적위험감소율
심장마비	11.8%	8.7%	26%
뇌졸중	5.7%	4.3%	25%
다른 모든 원인	14.7%	12.9%	12%

[HPS 임상시험 결과]
(The Art of Statistics, p103)

스타틴은 심장마비와 뇌졸중을 예방하는데 도움을 주는가?

- 임상시험 결과 심장마비의 경우 절대 위험도가 11.8-8.7=3.1% 감소되었다. 즉 스타틴 복용 시 1,000명 중 약 31명에게 심장마비 예방효과가 있었다.
- 그런데 임상시험 기간 동안 treatment group에서 18%가 스타틴 복용을 중단한 반면, control group에서 32%가 스타틴을 복용하기 시작했다. 여기서 제시된 결과는 intention to treat에 따라 보고되었기 때문에 실제 복용 효과는 제시된 결과보다 50% 더 높다고 추정한다.

스타틴은 심장마비와 뇌졸중을 예방하는데 도움을 주는가?

- 최근의 한 연구에서는 스타틴에 대한 27개의 RCT를 이용한 메타분석을 실시하였다.
- 이 연구에서 스타틴 효과는 LDL(나쁜 콜레스테롤) 감소에서 비롯되었다고 가정하고 스타틴의 복용 효과로 LDL이 1mmol/L 감소할 때마다 주요 혈관 질환의 발병가능성이 21%씩 감소한다고 결론을 내렸다.

오늘의 강의 요점

- 상관관계가 인과관계를 의미하지 않는다.
- 인과관계를 증명하기 위해서는 다음 2가지를 사용해야 한다.
 - 대규모 임상시험
 - 여러 번 반복 관측되는 결과

○ 출처

#1 <http://tylervigen.com/spurious-correlations>

#2 D. Spiegelhalter, (2019), The Art of Statistics, Penguin Random House

뭣이 중헌디? - 인과관계 알아보기

2. (관측연구에서) 인과관계를 보이려면?

기도는 효과가 있는가?

- 기도가 효과적인지 알아보기 위해 다음과 같은 관상동맥 우회술을 받은 환자 1,800명을 임의할당으로 3개의 그룹으로 나눈 후 다음과 같은 실험을 실시하였다.
 - 첫번째 그룹에서는 기도를 받았으나 환자들은 그 사실을 몰랐다.
 - 두번째 그룹에서는 기도를 받지 않았고 환자들은 그 사실을 몰랐다.
 - 세번째 그룹에서는 기도를 받았고 환자들은 그 사실을 알았다.
- 그런데 실험결과는 세번째 그룹에서 합병증에 시달리는 환자가 약간 증가했다!

왜 노인들은 귀가 클까?

- 노인들은 일반적으로 귀가 크다고 알려져 있다. 이 가설에 대한 검증을 위한 영국과 일본의 연구팀들이 횡단면연구 (cross-sectional study)를 통해서 자료를 수집해서 분석한 결과 명백한 상관관계가 있었다.
- 이 연관성은 어떻게 설명할 수 있을까? 다음과 같은 가설을 고려해 볼 수 있다.
 - 나이가 들면서 귀가 계속 자란다.
 - 귀가 작은 사람들은 일찍 죽는다.

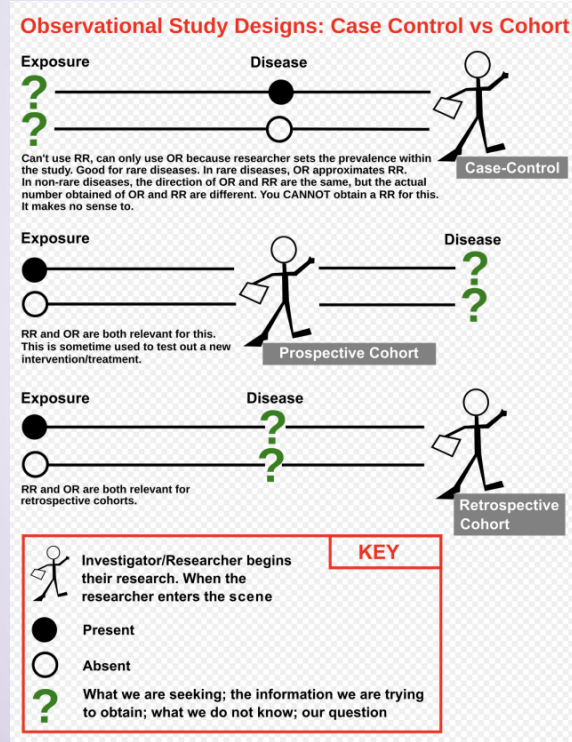
왜 노인들은 귀가 클까?

- 이러한 가설을 검증하기 위해서 다음과 같은 연구를 고려할 수 있다.
- 젊은이들을 생애별로 추적조사하여 귀의 크기 변화를 측정하거나 또는 작은 귀를 가진 사람이 더 빨리 죽는지 확인한다. 이러한 연구를 전향적 코호트 연구 (prospective cohort study)라고 한다.

왜 노인들은 귀가 클까?

- 전향적 코호트 연구는 시간이 오래 걸린다는 단점이 있다. 대안으로 후향적 코호트 연구(retrospective cohort study)를 고려할 수 있다. 즉 현시점에서 나이가 많은 사람들을 골라서 과거의 사진을 골라 그들의 귀가 커졌는지를 고려하는 것이다.
- 코호트 연구와 대조가 되는 연구로 사례-대조 연구(case-control study)를 들 수 있다. 이 경우 나이와 그 밖에 장수 여부에 영향을 미치는 요인이 서로 비슷한 사망자와 생존자 사이의 귀의 크기를 비교하는 것이다.

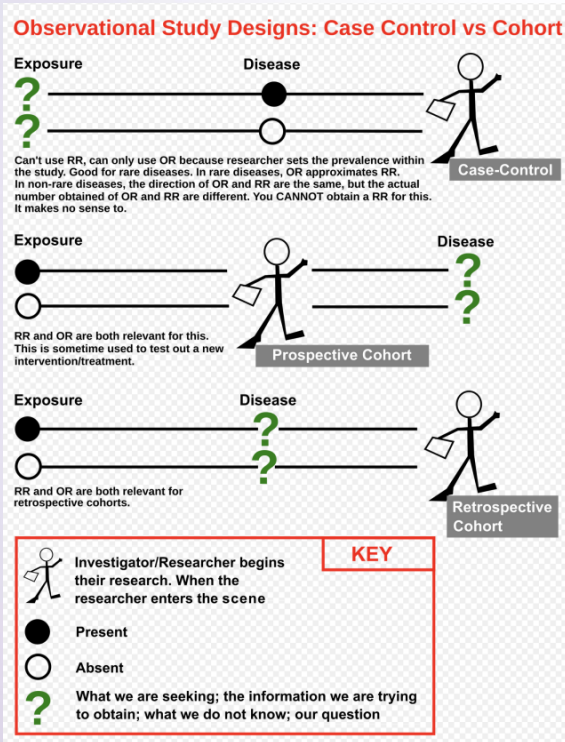
코호트 연구 vs 사례-대조군 연구



- 관측연구에서 코호트 연구와 사례-대조군 연구의 차이는 비교집단을 나누는 기준이 무엇인지에 달려있다.
- 사례-대조군연구에서는 비교집단을 질병유무로, 코호트 연구에서는 위험요인에 노출되었는지 유무에 따라 그룹을 나눈다.

[사례 대조군 연구 vs 코호트 연구]
(Wikipedia)

코호트 연구 vs 사례-대조군 연구



- 앞의 귀의 크기와 수명에 관한 연구의 경우 귀의 크기를 위험 요인(risk factor), 사망여부를 질병유무(outcome)으로 생각할 수 있다.

[사례 대조군 연구 vs 코호트 연구]
 (Wikipedia)

중첩요인 (confounder)

- 관측연구에서 인과관계를 설명하기는 매우 힘들다. 예를 들면 아이스크림 판매와 익사 간의 높은 상관관계는 사실 더운 날씨라는 숨어 있는 요인에 영향을 받은 것으로 생각할 수 있다.
- 이렇게 연관성을 보이는 두 결과에 동시에 영향을 미치는 공통 요인을 중첩요인 (confounder)라고 한다.

중첩요인 (confounder)

- 심슨의 역설이 중첩요인을 설명하는 좋은 예라고 할 수 있다. 버클리 대학의 대학원 합격율과 지원자의 성별 사이의 관계는 전공이라는 중첩요인으로 때문이었다.
- 중첩요인의 영향을 배제하기 위해서 중첩요인의 값이 같다는 가정하에서 원하는 두 변수사이의 관계를 알아보면 된다. 즉 층화(stratification)를 하거나 혹은 다음에 배울 회귀분석을 사용하여 중첩요인을 통제(control)하는 것이다.

건물에 스타벅스가 입점하면 집값이 오른다

- 스타벅스가 입점하면 주변 집값이 오른다는 얘기가 있다.
- 하지만 스타벅스의 경우 입점을 하기 위해 주변 상권을 면밀히 분석한 후 입점 여부를 결정한다는 점을 간과해서는 안된다.
- 즉 스타벅스때문에 집값이 올라가는 것이 아니라 (상권이 발달가능성이 높아서) 집값이 오를 가능성이 있는 곳에 스타벅스가 입점해서 나타나는 현상일 수 있다.
- 이처럼 실제 주장한 인과관계와 정확히 상반되는 관계를 역인과관계(reverse causation)라고 한다.

적당히 술을 마시면 건강에 좋다?

- 음주와 건강 간의 연관성 연구에서 술을 적당히 마시는 사람들이 전혀 마시지 않는 사람보다 사망률이 낮다고 알려져 있다.
- 이 연구의 결과는 역인과관계에 기인한 것으로 간주할 수 있다. 즉 건강이 너무 좋지 않은 사람들은 술을 마시지 못하기 때문에 음주자 집단과 비음주자 집단에 속한 사람들의 건강상태가 비슷하지 않은 경우가 있을 수 있다.

관측연구에서 인과관계를 도출하려면?

- 영국의 통계학자 오스틴 힐은 최초의 임상시험을 설계하였고 이 후 이 시험이 다른 RCT의 표준이 되었다.
- 1950년대에 오스틴 힐과 리처드 돌이 공동으로 흡연과 폐암관의 인과관계를 밝히는 연구를 이끌었고 연관성이 인과관계인지 판별하기 위해 Hill's criteria가 만들어 졌다. 참고로 지금 제시하는 Hill's criteria는 다른 학자들에 의해 조금 변형된 버전이다.
- Hill's criteria는 크게 직접적 증거, 메커니즘 증거, 평행증거로 나누어 진다.

직접적 증거

- 효과의 크기: 효과의 크기가 너무 커서 중첩요인으로 설명할 수 없다.
- 적절한 시공간적 근접성: 원인과 결과가 밀접한 시공간상에서 관측된다.
- 용량반응성(dose response)과 가역성(reversibility): 위험요인이 증가하면 그 효과도 따라서 증가하고 위험요인이 감소할 경우 효과도 감소한다.



메커니즘 증거

- 인과고리를 설명해 줄 그럴듯한 생물학적, 화학적, 기계적 메커니즘과 외적증거가 존재한다.

평행 증거

- 그 효과가 기존의 사실과 잘 들어 맞는다.
- 동일한 효과가 해당 연구를 재현했을 때 발견된다.
- 동일한 효과가 유사 연구에서 발견된다.

오늘의 강의 요점

○ 관측연구의 3가지 유형

- 코호트 연구
- 사례-대조군 연구
- 횡단면 연구

○ 관측연구에서 인과관계를 도출하려면?

- 직접적인 증거
- 메커니즘 증거
- 평행 증거

○ 출처

#1 Wikipedia https://en.wikipedia.org/wiki/Retrospective_cohort_study

뭣이 중헌디? - 인과관계 알아보기

3. 중첩요인을 통제하려면?

회귀분석을 이용한 중첩요인 통제

- 앞의 강의에서 중첩요인이 있을 경우 이를 통제하는 방법은 총화와 가중평균에 대해서 알아보았다.
- 보다 일반적인 중첩요인을 통제하는 방법으로 다중회귀분석을 들 수 있다.
- 회귀분석을 소개하기 위해 우리는 다윈의 사촌 골턴에게로 다시 잠시 돌아가자.
- 골턴은 부모의 키를 이용해서 자녀의 키를 예측하는데 관심이 있었다.

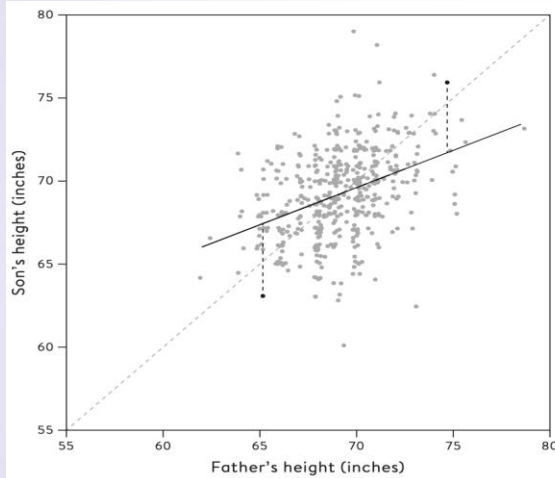
아버지의 키로 아들의 키를 예측할 수 있을까?

- 1886년 골턴은 상당수의 부모와 자녀에 관한 키 자료를 수집하였다. 이 자료의 요약치는 다음과 같다.

	인원	평균	중앙값	표준편차
어머니	197	64.0	64	2.4
아버지	197	69.3	69.5	2.6
딸	433	64.1	64.0	2.4
아들	465	69.2	69.2	2.6

[아버지와 아들의 키]
(The Art of Statistics, p164)

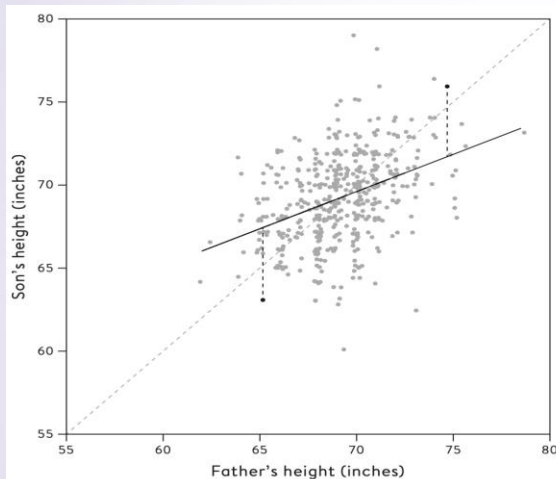
아버지의 키로 아들의 키를 예측할 수 있을까?



[아버지와 아들의 키]
(The Art of Statistics, p164)

- 왼쪽 그림은 아버지 (195명)와 아들(465명)의 키의 산점도이다. 여기서 피어슨 상관계수의 값은 0.39이다.

아버지의 키로 아들의 키를 예측할 수 있을까?



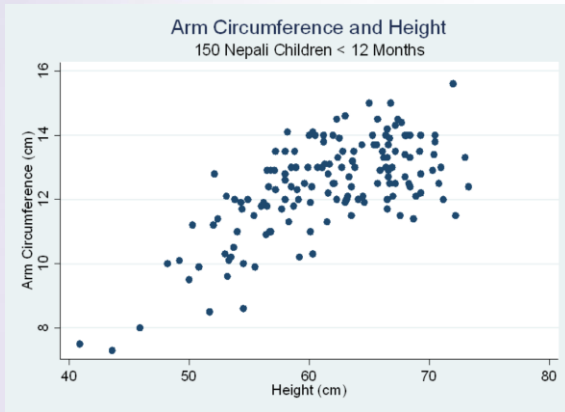
[아버지와 아들의 키]
(The Art of Statistics, p164)

- 만약 우리가 아버지와 아들의 키를 관계를 나타내는 임의의 직선을 그린다면 그 직선을 이용한 아들의 키의 예측치와 실제 아들의 키와의 차이를 잔차 (residual)이라고 한다.
- 이러한 잔차들의 차이를 최소화하는 방법으로 최소제곱법 (least-square method)을 사용할 수 있다.

유아의 팔목 두께가 굵으면 키도 크다?

- 우리는 팔목 두께와 키와의 연관성이 있는지 알고 싶다.
- 12개월 미만의 네팔 유아 150명의 신체 지수에 관한 자료를 이용하여 위의 가설에 검증하자. 이 자료의 요약치는 다음과 같다.
- 키: 평균 61.6cm, 표준편차 6.3cm,
범위 40.9cm~73.3cm
- 팔목두께: 평균 12.4cm, 표준편차 1.5cm,
범위 7.3cm~15.6cm

유아의 팔목 두께가 굵으면 키도 크다?

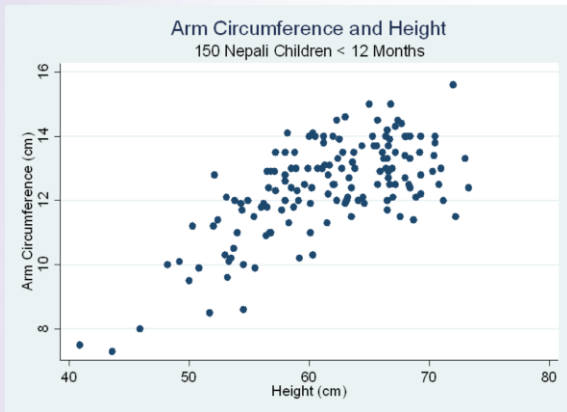


[McGready 교수의 Statistical Reasoning II 강의노트]
(JHSPHOPEN)

- 팔목두께를 키로 설명하는 회귀직선을 적합시킨 결과 추정식은 다음과 같다. 여기서 x_1 는 키, \hat{y} 는 팔목두께 추정치이다.

$$\hat{y} = 2.7 + 1.6 \cdot x_1$$

유아의 팔목 두께가 굵으면 키도 크다?



[McGready 교수의 Statistical Reasoning II 강의노트]
(JHSPHOPEN)

○ 기울기추정치 ($=0.16$)에 대해서는 두가지 해석이 가능하다.

- 유아의 키가 1cm 증가 시 팔목 두께는 평균 0.16cm 증가한다.
- 키 차이가 1cm인 유아들을 두 그룹으로 나누었을 경우 두 그룹의 팔목 두께의 차이의 평균이 0.16cm(키가 큰 그룹이 팔목 두께가 0.16cm 굵다.)

유아의 팔목 두께, 키 몸무게의 관계는?

- 만약 우리가 예측변수로 유아의 몸무게를 포함하면 어떻게 될까? 이 경우 회귀직선은 다음과 같이 주어진다. 여기서 x_2 는 유아의 몸무게를 나타낸다.

$$\hat{y} = 7.8 + 0.8 \cdot x_2$$

- 이제 예측변수로 유아의 키와 몸무게를 동시에 포함해보자. 이 경우의 회귀직선은 다음과 같다.

$$\hat{y} = 14.1 - 0.16 \cdot \text{height} + 1.40 \cdot \text{weight}$$

- 키의 기울기가 양수에서 음수로 바뀌었다!

유아의 팔목 두께, 키 몸무게의 관계는?

- 단순회귀모형과 다중회귀모형을 결과를 비교하면 다음과 같다.

	키만 예측변수인 경우	몸무게만 예측변수인 경우	둘다 예측변수인 경우
키	0.16		-0.16
몸무게		0.80	1.40

[단순회귀모형과 다중회귀모형의 비교]

유아의 팔목 두께, 키 몸무게의 관계는?

- 이 모형에서 키의 기울기는 -0.16 으로 다음과 같이 해석할 수 있다. 몸무게가 같고 키가 1cm 차이나는 두 유아그룹들간의 평균 팔목 두께 차이는 0.16cm 이며 키가 1cm 큰 그룹이 팔목 두께가 평균적으로 0.16cm 작다.
- 단순회귀분석과 달리 이 기울기는 다른 변수(몸무게)의 연관성을 보정한 추정치다. 즉 몸무게가 같은 두 아이의 키 차이가 난다면 작은 아이는 통통한 편이라는 것을 짐작할 수 있기 때문에 사실 이 경우 키와 팔목 두께는 음의 연관성을 가지는 것이 당연하다!

오늘의 강의 요점

- 중첩요인을 통제하는 방법으로 층화, 가중평균, 회귀분석을 사용할 수 있다.

○ 출처

#1~2 D. Spiegelhalter, (2019), The Art of Statistics, Penguin Random House

#3 JHSPHOPEN

<http://ocw.jhsph.edu/index.cfm/go/viewCourse/course/StatisticalReasoning2/coursePage/lectureNotes/>

뭇이 중헌디? - 인과관계 알아보기

Lab 5: R과 RStudio 사용법 소개

R 소개

- R은 통계 계산에 적합한 프로그래밍 언어로서 MATLAB과 같은 다른 과학 계산을 언어와 비교 시 많은 장점이 있다.
- 이미 다른 프로그래밍 언어에 익숙한 학생들은 큰 어려움 없이 배울 수 있으며 다양한 무료 참고 문헌을 다음 웹사이트에서 찾을 수 있다. <http://www.r-project.org/>

R 소개


- R에 관한 일반적인 소개는 다음 뉴욕 타임스 기사를 참조하기 바란다 .http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=0
- R에 관한 무료 한글 참고 문헌을 원하는 학생들은 서민구 (2014) R을 이용한 데이터 실무분석 (<http://r4pda.co.kr/>)을 권한다.

R install하기

- 먼저 <https://cran.r-project.org/>로 간 뒤 본인의 컴퓨터 운영체제용 R 프로그램을 선택한다. 국내 컴퓨터 사용자의 절대다수가 윈도우를 사용하기 때문에 윈도우를 기준으로 설명하면 다음과 같은 화면을 볼 수 있다. 여기서 R version은 2020년 10월 10일 기준으로 4.0.3이며 계속 변화할 수 있다는 점을 기억하자.

R 설치하기





CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

R-4.0.3 for Windows (32/64 bit)

[Download R 4.0.3 for Windows](#) (85 megabytes, 32/64 bit)
[Installation and other instructions](#)
[New features in this version](#)

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [md5sum](#) of the .exe to the [fingerprint](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages in my previous version of R?](#)
- [Should I run 32-bit or 64-bit R?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is CRAN.MIRROR>/bin/windows/base/release.html.

Last change: 2020-10-10

[R-4.0.3 for Windows(32/64 bit)]
(R project)

R install하기



RGui (64-bit)

File Edit View Misc Packages Windows Help

R Console

```
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.


> |
```

[R-4.0.3 for Windows(32/64 bit)]

RStudio란?


RStudio Desktop 1.3.1093 - [Release Notes](#)

1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:



DOWNLOAD RSTUDIO FOR WINDOWS
1.3.1093 | 171.62MB

Requires Windows 10/8/7 (64-bit)



All Installers

Linux users may need to [import RStudio's public code-signing key](#) prior to installation, depending on the operating system's security policy.

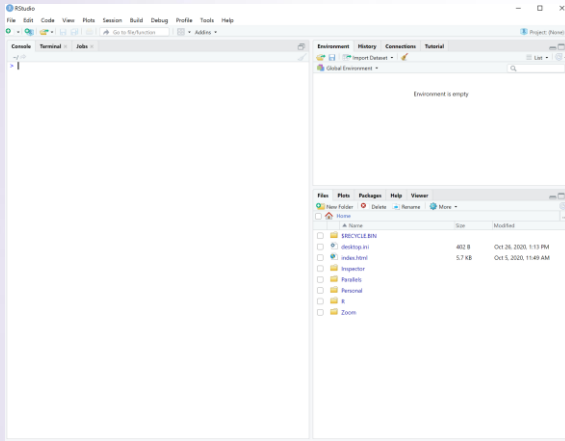
RStudio requires a 64-bit operating system. If you are on a 32 bit system, you can use an [older version of RStudio](#).

OS	Download	Size	SHA-256
Windows 10/8/7	RStudio-1.3.1093.exe	171.62 MB	62b9e60a
macOS 10.13+	RStudio-1.3.1093.dmg	148.66 MB	bd0c4d3a4

[Rstudio-1.3.1093 for Windows]
(R Studio)

- R을 사용하기 위해 통합개발 환경 (IDE) 프로그램으로 무료버전 RStudio Desktop을 다운로드받을 수 있다. (<https://rstudio.com/products/rstudio/download/>)

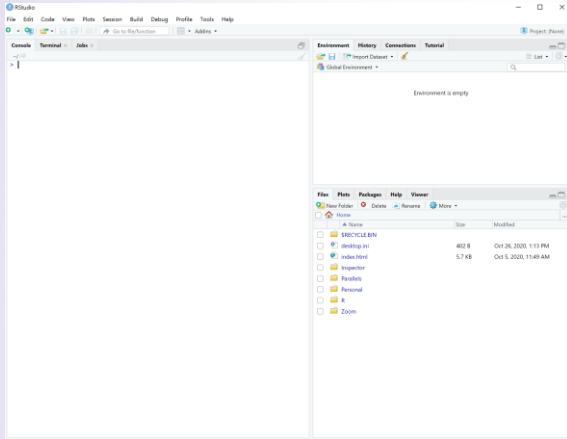
RStudio 설치하기



[Rstudio 시작화면]

- RStudio 을 설 치 한 후 RStudio프로그램을 열어보면 다음과 같은 화면을 볼 수 있다.
- RStudio의 화면은 여러 개의 구획으로 이루어져 있다. 일반적으로 보이는 구획은 다음과 같다.

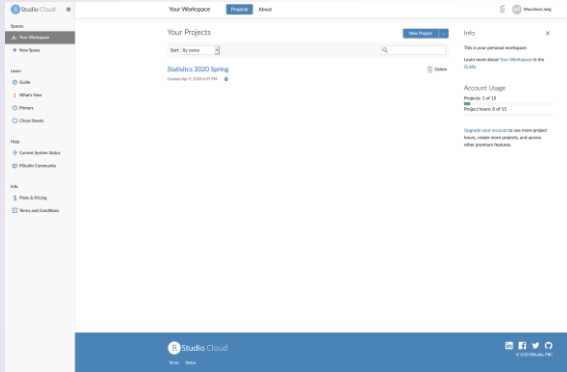
RStudio 설치하기



[Rstudio 시작화면]

- 콘솔: R 명령을 입력하고 결과가 출력된다.
- 환경: 현재 세션에서 사용 가능한 각 객체를 보여준다.
- 파일: 디렉터리의 파일을 보여준다.

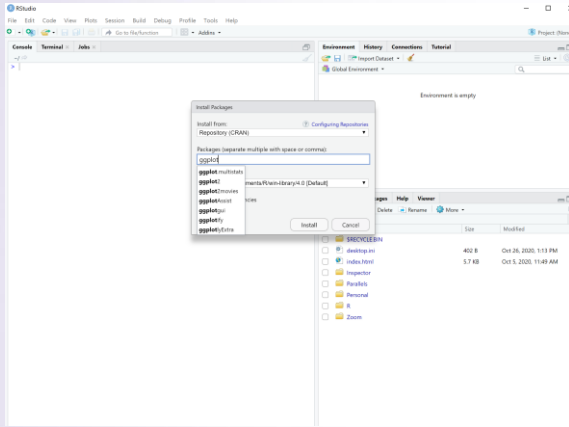
프로그램 설치가 힘들다면



[Rstudio Cloud 시작화면]

- 프로그램을 인스톨하지 않게 웹 기반으로 사용을 원할 경우 RStudio에서 운영하는 클라우드 기반 서비스 RStudio Cloud (<https://rstudio.cloud/>)를 사용할 수 있다.

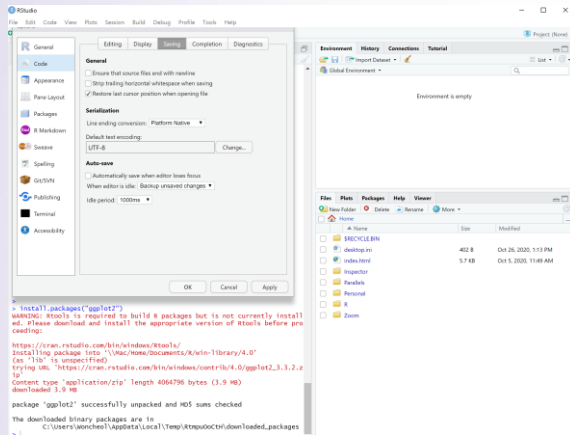
R Package 설치



[R Package 설치]

- R의 가장 큰 장점은 다양한 분석 기능을 가진 수많은 Package를 제공한다는 점에 있다.
- RStudio menu에서 Tools > Install Packages를 선택해서 대화상자를 사용하여 “ggplot2”라는 package를 설치해보자.

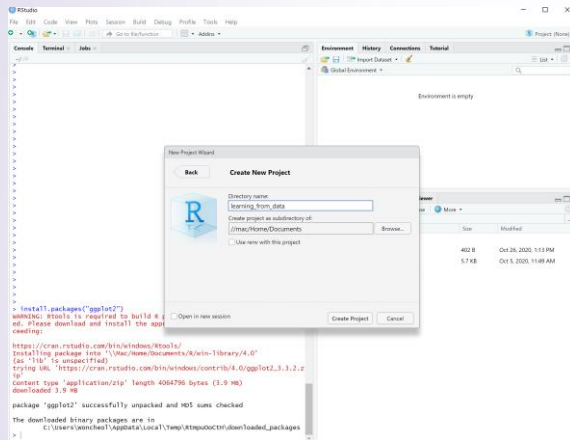
한글을 보려면...



[Encoding 변환]

○ RStudio 에서 Tools > Global Options > Code > Saving 을 선택한 후 Defaulting text encoding 을 UTF-8 으로 변경하고 Apply 버튼 클릭

한글을 보려면...



[Working directory 설정]

모든 분석은 프로젝트별로 별도의 디렉터리에 만들어서 관리하는 것이 편하다.

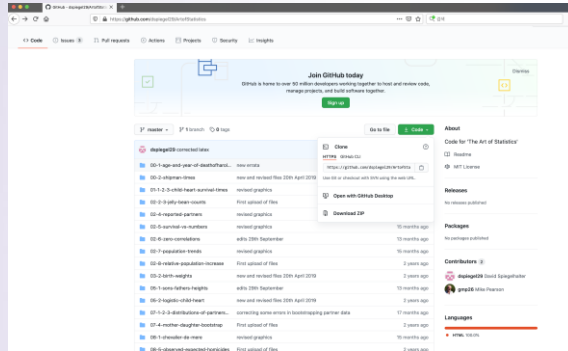
File > New Project > New Directory > New Project

대화상자에서

프로젝트-디렉터리 이름: learning_from_data

상위 디렉터리: browsing을 통해서 적절한 디렉터리 선정

데이터 읽기



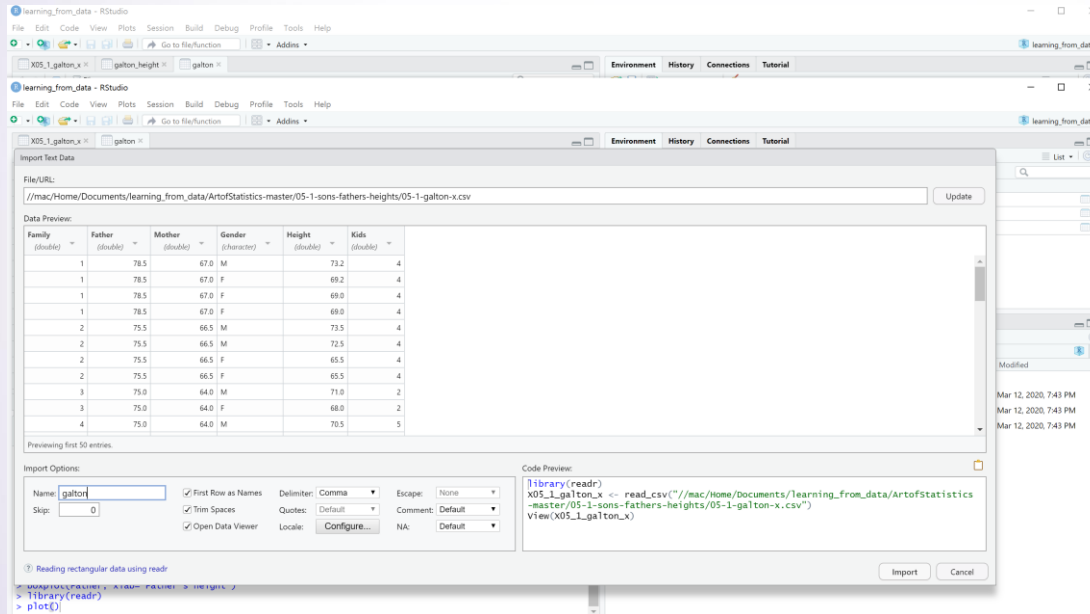
[Art of Statistics Github 페이지]
(Github 홈페이지)

- The Art of Statistics에 사용된 모든 자료와 R code는 책의 Github 페이지 (<https://github.com/dspiegel29/ArtofStatistics>)에서 다운로드 받을 수 있다.
- 화면에서 code라고 표시되어 있는 초록색 버튼을 클릭한 후 “Download ZIP”을 클릭하면 전체 파일을 하나의 압축 파일로 다운로드 받을 수 있다.

자료 읽기

- 다운로드받은 압축 파일을 앞의 프로젝트 디렉터리에서 풀면 “ArtofStatistics-master” 라는 sub-directory 가 생긴다.
- 그 아래의 05-1-sons-fathers-heights 라는 sub-directory 아래 05-1-galton-x.csv을 클릭할 경우 2가지 메뉴가 나오는데 “Import Datset..”이라는 메뉴를 클릭하자. 이 경우 필요한 R library “readr”, “Rcpp”을 설치할 지 여부를 물어보면 yes라고 답하면 된다.
- 이 후 나타나는 창 왼쪽 아래의 데이터 셋의 이름을 물어보는 창에 “galton”라고 쓰고 오른쪽 아래의 Import 버튼을 누르면 골턴의 부모와 자녀 키 자료를 읽게 된다.

자료 읽기



learning_from_data - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Environment History Connections Tutorial

learning_from_data

learning_from_data - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Environment History Connections Tutorial

Import Text Data

File/URL: //mac/Home/Documents/learning_from_data/ArtofStatistics-master/05-1-sons-fathers-heights/05-1-galton-x.csv Update

Data Preview:

Family (double)	Father (double)	Mother (double)	Gender (character)	Height (double)	Kids (double)
1	78.5	67.0	M	73.2	4
1	78.5	67.0	F	69.2	4
1	78.5	67.0	F	69.0	4
1	78.5	67.0	F	69.0	4
2	75.5	66.5	M	73.5	4
2	75.5	66.5	M	72.5	4
2	75.5	66.5	F	65.5	4
2	75.5	66.5	F	65.5	4
3	75.0	64.0	M	71.0	2
3	75.0	64.0	F	68.0	2
4	75.0	64.0	M	70.5	5

Previewing first 50 entries.

Import Options:

Name: galton

Skip: 0

☒ First Row as Names

☒ Trim Spaces

☒ Open Data Viewer

Delimiter: Comma

Quotes: Default

Locale: Configure...

Escape: None

Comment: Default

NA: Default

Code Preview:

```
library(readr)
X05_1_galton_x <- read_csv("//mac/Home/Documents/learning_from_data/ArtofStatistics-master/05-1-sons-fathers-heights/05-1-galton-x.csv")
View(X05_1_galton_x)
```

Import Cancel

Reading rectangular data using readr

```
> library(readr)
> plot()
```

[골턴의 키 자료 읽기]

골턴 자료의 요약치

- Art of Statistics 의 Github페이지에 이 자료 분석에 사용된 R code가 있다 (<https://bit.ly/3jLZ67c>).

```
galton<-read.csv("05-1-galton-x.csv",header=TRUE) # read csv file into dataframe galton
attach(galton) #uncomment if/while necessary

summary(galton)
```

```
##      Family      Father      Mother      Gender      Height
## 185      : 15   Min.    :62.00   Min.    :58.00   F:433   Min.    :56.00
## 166      : 11   1st Qu.:68.00   1st Qu.:63.00   M:465   1st Qu.:64.00
## 66       : 11   Median :69.00   Median :64.00           Median :66.50
## 130      : 10   Mean    :69.23   Mean    :64.08           Mean    :66.76
## 136      : 10   3rd Qu.:71.00   3rd Qu.:65.50           3rd Qu.:69.70
## 140      : 10   Max.    :78.50   Max.    :70.50           Max.    :79.00
## (Other):831
##      Kids
## Min.    : 1.000
## 1st Qu.: 4.000
## Median : 6.000
## Mean    : 6.136
## 3rd Qu.: 8.000
## Max.    :15.000
##
```

[골턴의 키 자료의 Summary]
(The Art of Statistics, p124)



오늘의 강의 요점

- R과 RStudio install 하는 법
- R을 사용한 간단한 자료분석

○ 출처

#1 <https://cran.r-project.org/>

#2 Copyright 2020. 장원철 all right reserved

#3 <https://rstudio.com/products/rstudio/download/>

#4~8 Copyright 2020. 장원철 all right reserved

#9 <https://github.com/dspiegel29/ArtofStatistics>

#10 Copyright 2020. 장원철 all right reserved

#11 D. Spiegelhalter, (2019), The Art of Statistics, Penguin Random House