데이터로 배우는 통계학

자연과학대학 통계학과
장원철 교수

우리 집 가격은 얼마지? – 회귀모형

1. 최소제곱법

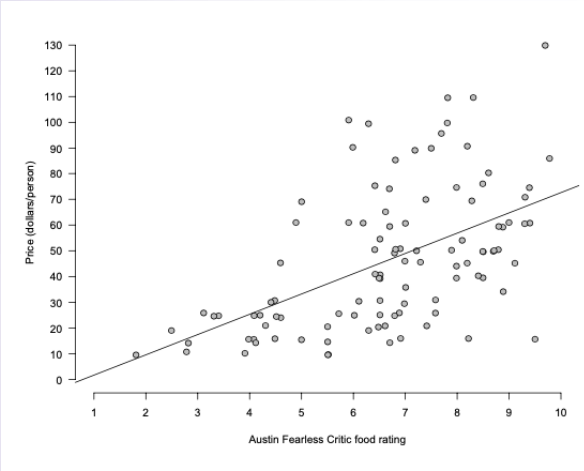
예측, 예측, 예측

- 지난 선거 결과를 바탕으로 이번 총선에서 각 정당의 예상 의석 수를 예측할 수 있을까?
- 집 가격을 복덕방을 통해서가 아니라 데이터를 이용하여 추정할 수 있을까?
- 최저임금 인상이 고용률에 어떤 영향을 주었을까?

설명변수와 반응변수

- 회귀분석은 두 변수 간의 관계를 모형화하는 방법이며 특히 한 변수를 이용하여 다른 변수를 예측하거나 설명하는 데 유용하다.
- 여기서 반응변수(response variable)는 우리가 설명하거나 예측하고 싶은 변수를 말하며 종속변수(dependent variable)라고도 불린다.
- 설명변수(explanatory variable)는 반응변수의 값을 예측하기 위해 사용되는 변수로 독립변수(independent variable)라고도 한다.

음식평가와 가격



[음식평가점수 vs 가격 Scott (2020)]
(Data Science, p32)

- 그림은 미국 텍사스주의 주도인 오스틴 중심가 음식점의 가격과 음식평가 점수와의 산점도를 보여준다.
- 이 두 변수의 관계를 나타내는 직선은

$$y = -6.2 + 7.9x$$이다.

기타 출처 #1

회귀분석 모형

- 앞의 예제와 같이 두 변수 사이의 관계가 선형관계라고 생각된다면 다음과 같은 식으로 각 각의 관측치(x_i, y_i)의 관계를 나타낼 수 있다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- 여기서 β_0는 절편, β_1은 기울기를 나타내며 ϵ_i를 오차항이라고 부른다.

- 그렇다면 우리는 어떻게 두 변수의 관계를 나타내는 위의 식을 추정할 수 있을까?

회귀분석 모형

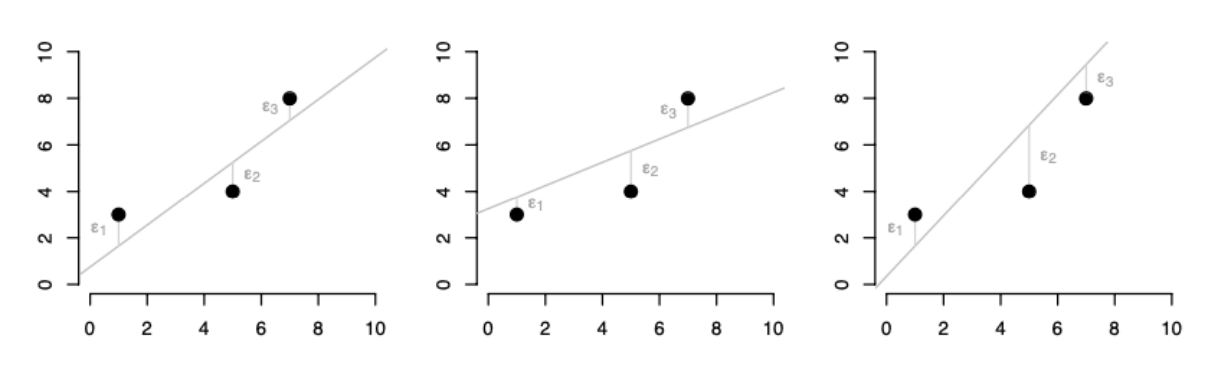
- 회귀분석 모형의 추정식은 다음과 같은 식으로 표현한다.

$$\hat{y} = b_0 + b_1 x$$

여기서 $\hat{y}$은 y의 예측치(즉 음식평가값 x이 주어졌을 경우 가격의 예측치), b_0와 b_1은 각각 절편과 기울기의 추정치이다.

- 추정치와 실제값의 차이를 잔차(residual)이라고 하는데 오차항의 추정치라고 할 수 있다. 즉 잔차는 추정치와 실제값과 수직거리 차이이다.

회귀모형의 추정



[음식평가점수 vs 가격 Scott (2020)]
(Data Science, p34)

- 만약 3개의 점이 주어진 경우 거기에 적합할 수 있는 최선의 회귀직선은 무엇일까?

기타 출처 #2

최소제곱법

- 최소제곱법은 잔차의 제곱 합을 최소로 하는 기울기와 절편의 값을 구하는 것이다. 즉 아래 식을 최소로 하는 b_0와 b_1를 찾는 것이다.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i)^2$$

- 위의 식을 최소로 하는 b_0와 b_1은 다음과 같다.

$$b_1 = r\frac{SD_y}{SD_x}, b_0 = \overline{y} - b_1\overline{x}$$

- 여기서 r는 피어슨 상관계수이며 SD_y는 y의 표준편차, SD_x는 x의 표준편차이다.

최소제곱법의 유래

- 최소제곱법은 프랑스 수학자 아드리앵-마리 르장드르가 1805년에 발표한 논문 “혜성 궤도를 결정하기 위한 새로운 방법”에서 처음 제시되었으나 독일 수학자 칼 프리드리히 가우스가 본인이 1795년부터 사용하고 있었던 방법이라고 1809년에 발표한 논문에서 주장하였다.

- 가우스는 최소제곱법을 확률이론과 정규분포를 연관하여 설명하였기 때문에 오늘날 최소제곱법의 창시자는 일반적으로 가우스로 간주한다.

오늘의 강의 요점

- 반응변수, 설명변수, 오차항
- 최소제곱법

우리 집 가격은 얼마지? – 회귀모형

2. 회귀모형의 진단

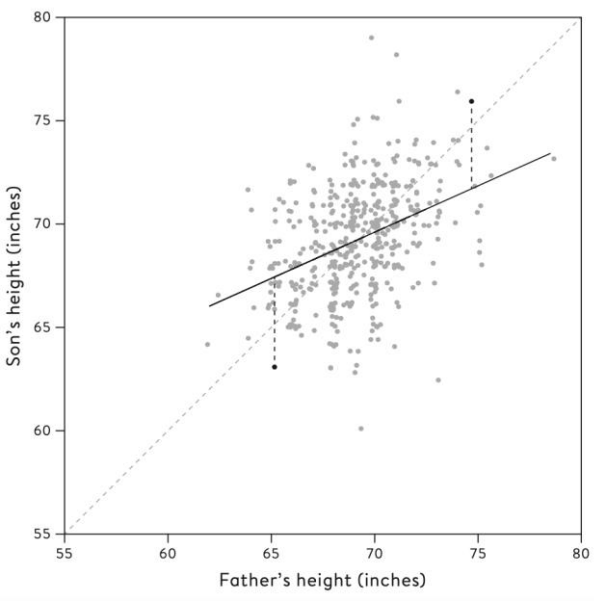
회귀모형

- 일반적인 회귀모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \dots, n$$

- 여기서 오차항 $\epsilon_1, \dots, \epsilon_n$ 은 서로 독립이고 평균이 0이고 분산이 σ^2 인 정규분포를 따른다고 가정한다.

회귀모형의 4가지 가정

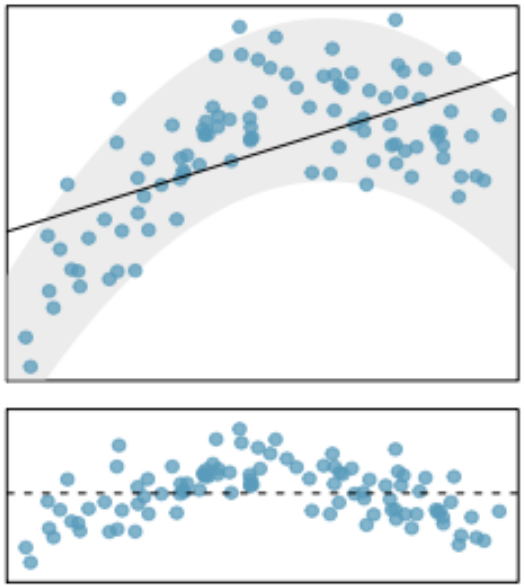


[아버지와 아들의 키]
(The Art of Statistics, p164)

- **회귀모형을 데이터에 적합시키기 위해서는 다음과 같은 가정이 필요하다.**
  - 반응변수와 설명변수간의 선형관계
  - 반응변수의 등분산성
  - 반응변수의 정규분포
  - 반응변수값의 독립

기타 출처 #3

선형관계

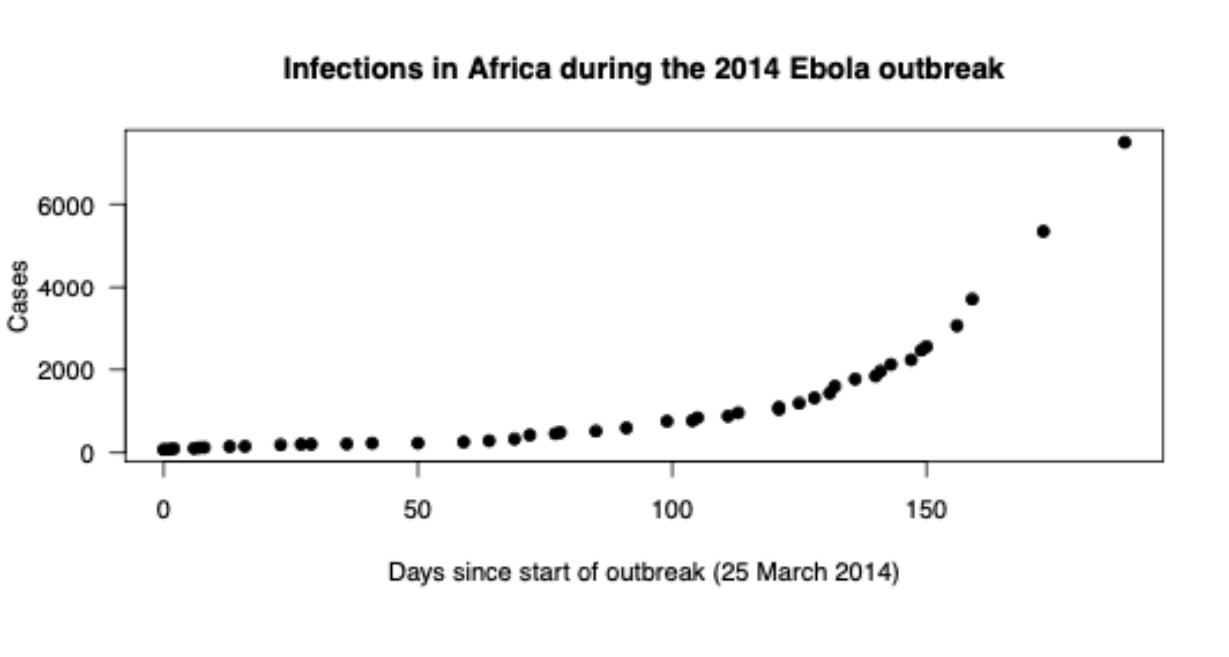


[선형관계]
(OpenIntro Statistics, p319)

- 반응변수와 설명변수관계가 선형일 경우 회귀분석을 사용할 수 있다.
- 선형여부에 관한 판단은 산점도 혹은 잔차 그림(residual plot)을 통해서 파악할 수 있다. 잔차그림은 설명변수와 설명변수에 대응하는 잔차와의 산점도를 의미한다.

기타 출처 #4

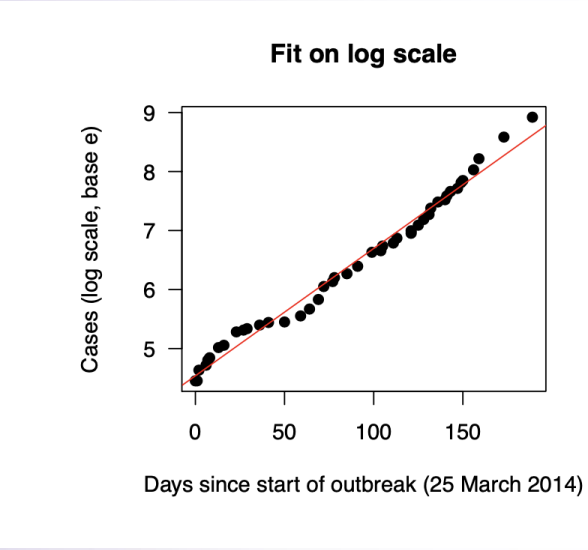
비선형 모형의 사례



[날짜 별 에볼라 발생건수, Scott, J. (2020)]
(Data Science, p48)

기타 출처 #5

전염병 확산모형



[날짜 별 log(에볼라) 발생건수, Scott, J. (2020)]
(Data Science, p49)

- 전염병 감염건수과 같이 기하급수적으로 증가하는 자료에 대해서는 반응변수에 로그 값을 취한 후 다시 회귀분석을 적합시킬 수 있다.
- 이 경우 적합된 회귀모형은

$$\log(cases) = 4.54 + 0.021 days$$

- 위의 식을 원래 데이터의 스케일로 바뀐다면

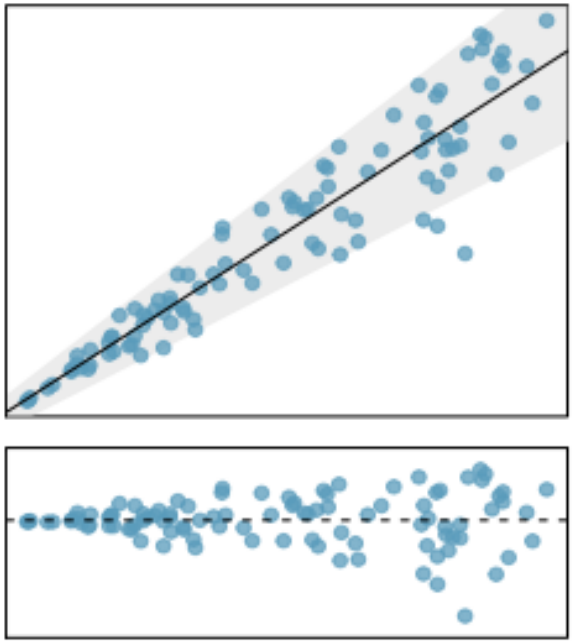
$$cases = 93.5 \cdot e^{0.021 \cdot days}$$

기타 출처 #6

로그변환하는 경우

- 반응변수의 값에 0이 포함된 경우 0에 아주 작은 숫자를 더한 후에 로그변환을 한다.
- 반응변수와 설명변수 모두 로그변환을 해야 하는 경우가 있다.

등분산성

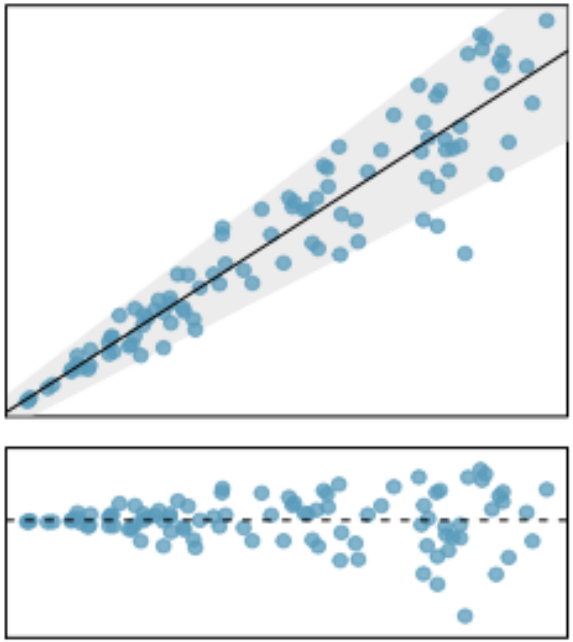


[등분산성]
(OpenIntro Statistics, p319)

기타 출처 #7

- 회귀모형에서 각 반응변수의 분산은 동일하다는 가정을 한다.
- 그림은 설명변수의 값이 증가함에 따라 반응변수의 분산이 증가하는 경향이 있음을 보여준다.

등분산성

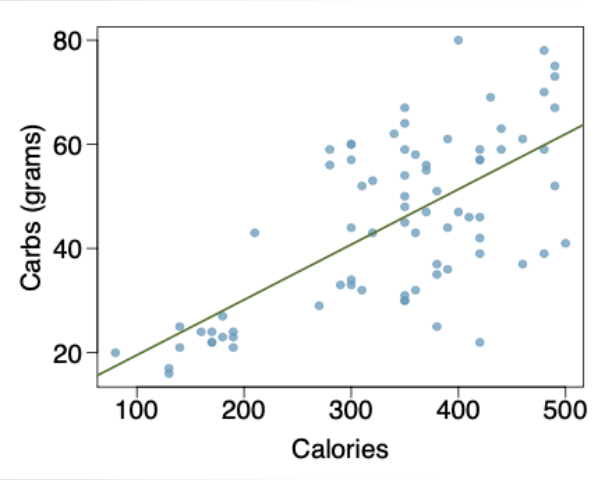


[등분산성]
(OpenIntro Statistics, p319)

기타 출처 #7

- 이 경우에 반응변수를 로그변환 혹은 제곱근 변환 등을 고려할 수 있다.
- 변환을 하지 않는다면 가중회귀모형 (weighted regress-ion model)을 고려할 수 있다.

정규성 가정

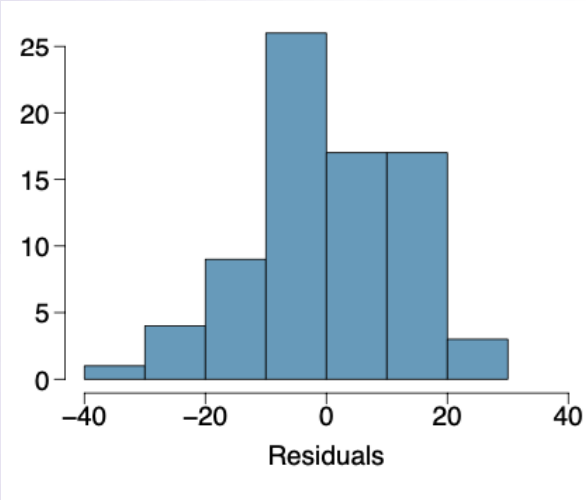


[Nutrition at Starbucks, Part I]
(OpenIntro Statistics, p326)

- 반응변수가 정규분포를 따른다는 것은 오차항이 정규분포를 따른다는 의미와 동일하다.
- 잔차들의 히스토 그램을 통해 정규성 가정을 확인할 수 있다.
- 그림은 스타벅스 메뉴 아이템의 칼로리와 탄수화물 함유량의 관계를 보여주고 있다.

기타 출처 #8

정규성 가정

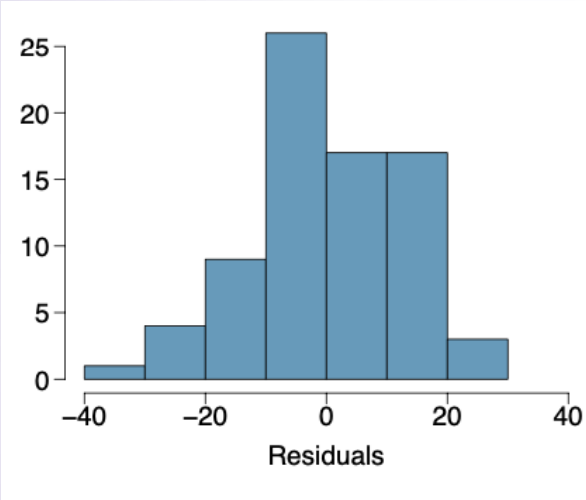


[Nutrition at Starbucks, Part Ⅰ]
(OpenIntro Statistics, p326)

- 왼쪽 잔차들의 히스토그램을 보면 분포가 대칭적이지 않아 정규분포와는 차이가 있음을 알 수 있다.
- 하지만 정규분포 조건 경우 아주 극단적으로 치우친 분포가 아니라면 회귀분석을 적합하여도 큰 문제는 없다!

기타 출처 #9

정규성 가정

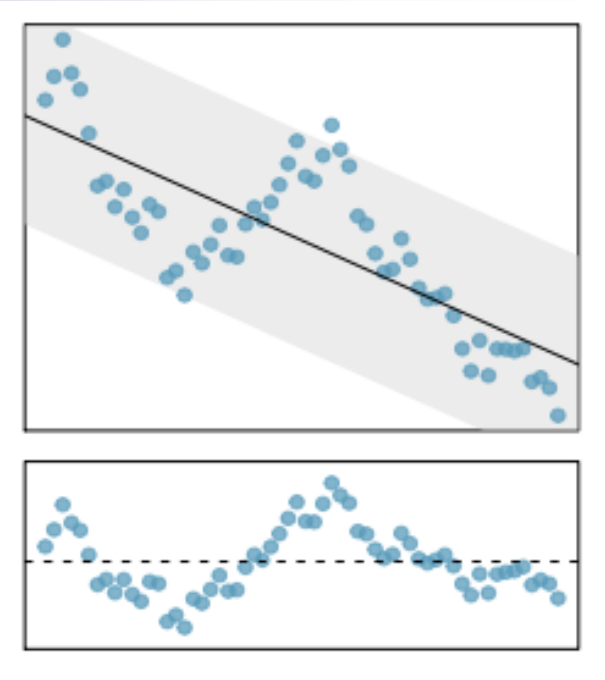


[Nutrition at Starbucks, Part Ⅰ]
(OpenIntro Statistics, p326)

- 이 예제의 경우 등분산성 가정을 위배해서 정규성 가정이 맞지 않는 것으로 볼 수 있기 때문에 등분산성 가정을 만족할 수 있는 변수 변환을 고려할 수 있다.

기타 출처 #9

독립성 가정



[독립성 가정]
(OpenIntro Statistics, p319)

- 반응변수의 값이 서로 의존하는 경우로는 주가지수와 같은 시계열 자료를 들 수 있다.
- 그림에서 연속되는 반응변수들의 값이 서로 강한 상관관계가 있음을 볼 수 있다.
- 이 경우 회귀모형을 사용할 수 없고 시계열 자료 분석을 위한 통계모형을 사용해야 한다.

기타 출처 #10

오늘의 강의 요점

- **회귀분석의 가정**
  - 선형관계
  - 등분산성
  - 정규성
  - 독립성

우리 집 가격은 얼마지? – 회귀모형

3. 회귀모형의 함정

중간고사를 잘 보면 기말고사는 못본다

- 중간고사를 잘 보면 기말고사는 (일반적으로) 못 본다.
- 신인상을 받은 선수가 그 다음해의 성적은 곤두박질 친다.
- 위의 사실들은 잘 알려진 사실이다. 골턴도 비슷한 현상을 아버지와 아들 키의 관계에서 발견했는데 아버지의 키가 평균보다 큰 경우 아들의 키는 아버지보다 일반적으로 작았으며 아버지의 키가 평균보다 큰 경우는 반대현상이 관측되었다.
- 골턴은 이러한 현상을 평범함으로의 회귀(regression to mediocrity)라고 했는데 오늘날은 평균으로의 회귀(regression to the mean)으로 알려져 있다.

과속 단속 카메라가 교통사고를 감소시키는가?

- 과속단속 카메라는 최근에 사고가 난 장소에 새로 설치가 된다.
- 설치 후에 사고율이 내려가면 사람들은 과속단속 카메라 때문이라고 믿는다. 사실일까?
- 하지만 평균으로의 회귀때문에 어차피 사고율은 떨어지게 되어있다!

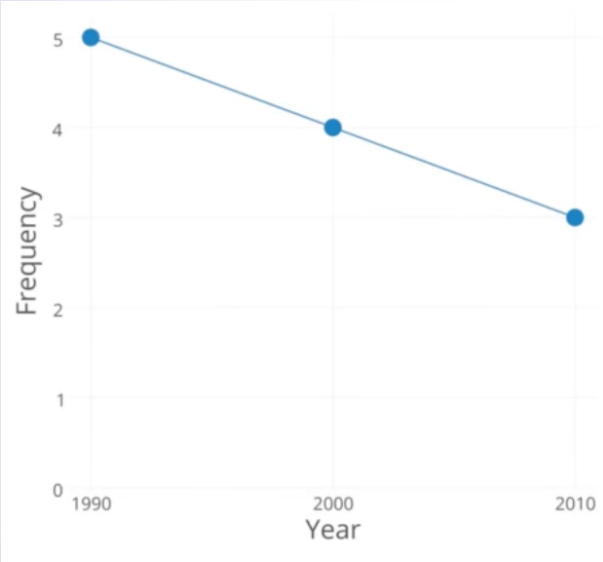
과속 단속 카메라가 교통사고를 감소시키는가?

- 국제학업성취도 비교연구(PISA)에 따르면 2003년 국가별 순위를 2012년 순위와 비교해보면 음의 상관관계를 가짐을 알 수 있다.
- 이 경우 피어슨 상관계수의 값은 -0.60인데 순위는 운으로 인한 것이고 순위변동이 평균으로의 회귀때문일때 상관계수의 값이 -0.71이라는 걸 고려한다면 국가간 순위는 사실상 거의 무의미하다고 할 수 있다.

과속 단속 카메라가 교통사고를 감소시키는가?

- 만약 정말 과속단속 카메라가 교통사고를 감소시키는 지 여부를 파악하려면 어떻게 해야 할 까?
- 카메라를 임의로 배치를 한 후 교통사고 감소 정도를 측정해야 한다. 실제 이런 실험을 한 결과 설치효과 중 2/3는 평균으로의 회귀로 인한 효과로 추정되었다.

영국인들은 2030년에는 더 이상 성관계를 하지 않는다?

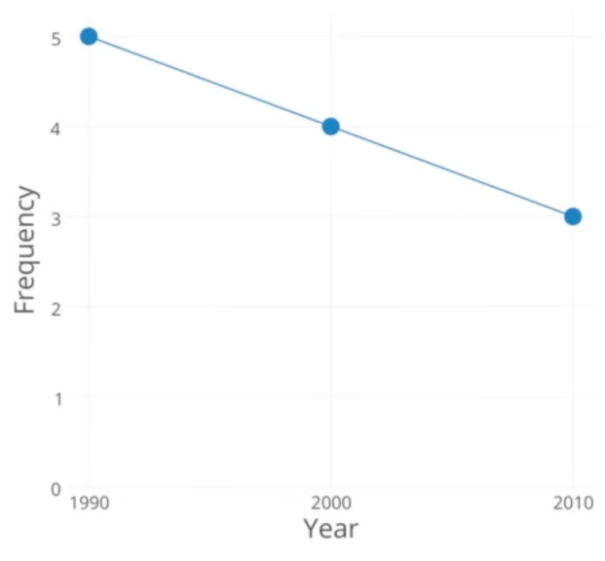


[Number of times the average person had sex in the past 4 weeks]
(Spiegelhalter's 2019 LES talk)

- 영국의 성생활 실태 설문조사에 따르면 최근 20년 동안 영국인의 월별 성관계 횟수는 40%정도 감소했다.

기타 출처 #11

영국인들은 2030년에는 더 이상 성관계를 하지 않는다?

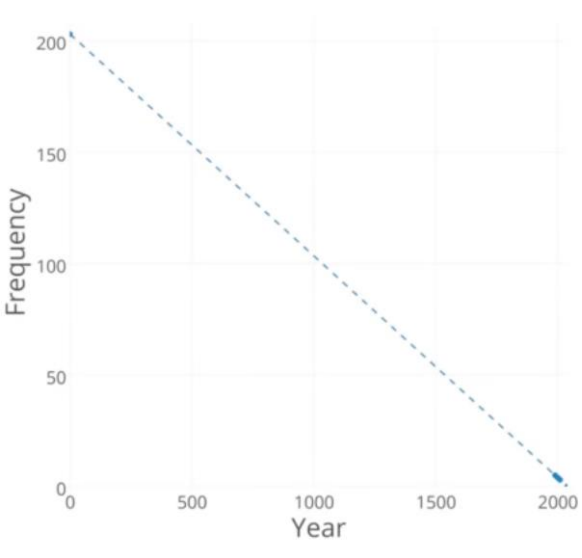


[Number of times the average person had sex in the past 4 weeks]
(Spiegelhalter's 2019 LES talk)

- 데이비드 스피겔헬터는 이 현상을 두고 2030년이 되면 영국인들은 더 이상 성관계를 하지 않을 지도 모른다는 (농담섞인) 예측을 하였다.
- 하지만 영국의 주요 일간지인 데일리 텔레그래프에서 이 예측(?)을 대대적으로보도하였다.

기타 출처 #11

외삽법 (Extrapolation)



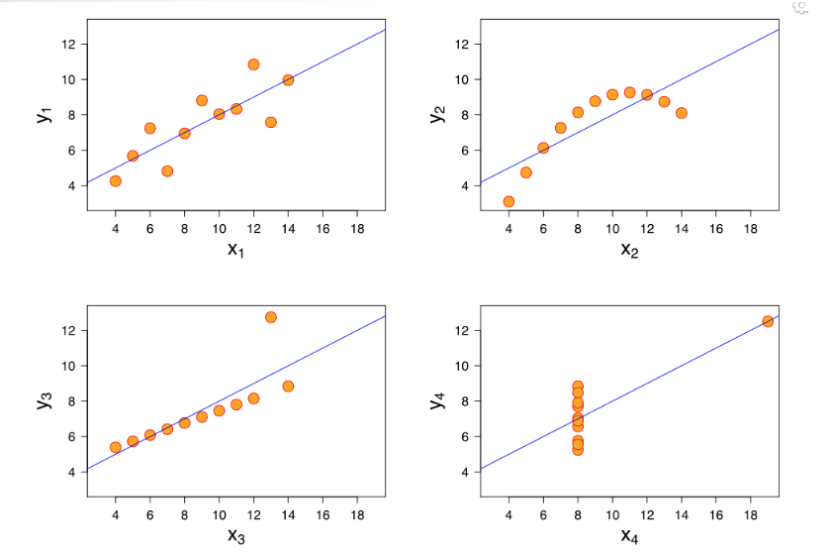
[Sex per month]
(Spiegelhalter's 2019 LES talk)

- 주어진 자료의 범위 밖에서 모형을 이용하여 예측하는 것을 외삽법 (exptraploation) 이라고 한다.
- 외삽법을 이용한 예측은 뉴스 매체에서 심심찮게 등장한다. 예를 들면 BBC는 2156년에는 여성이 남성을 달리기에서 앞지를 것이라는 예측을 보도하기도 했다.

기타 출처 #12

Anscombe's quartet

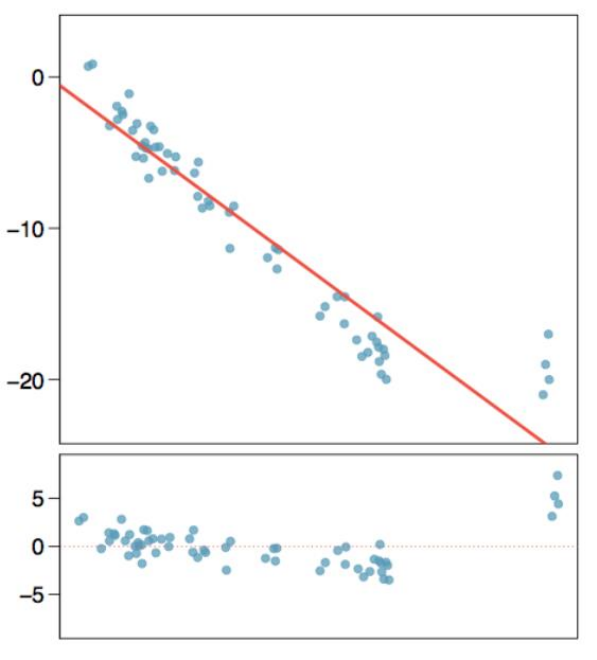
- 아래의 4개의 데이터 셋은 전혀 다른 형태를 보여주고 있지만 최소제곱법으로 계산한 회귀직선의 추정치는 모두 $\hat{y} = 3 + 0.5x$으로 같다!



기타 출처 #13

[안스콤의 퀴르텟]
(Wikimedia)

회귀분석에서 이상점

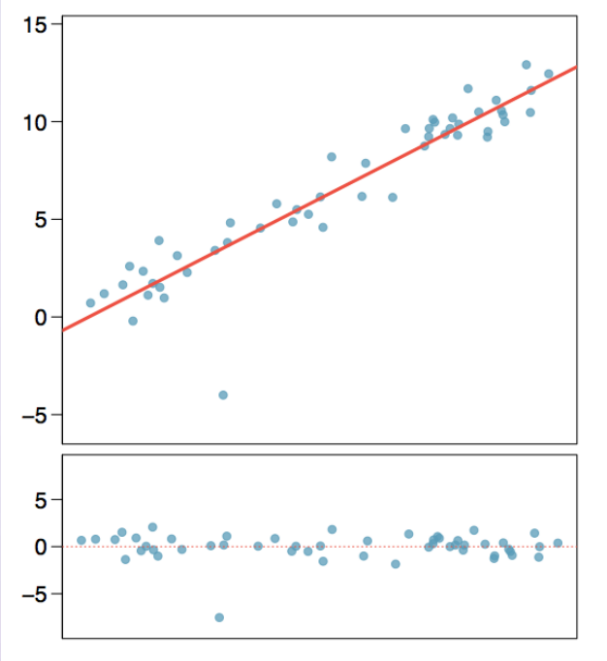


[회귀분석에서의 이상점]
(OpenIntro Statistics, p329)

- 왼쪽 위 그림은 데이터와 적합된 회귀직선, 그리고 아래 그림은 각 설명변수에서의 잔차값을 보여주고 있다.
- 만약 하단의 4점을 포함하지 않고 회귀직선을 구했다면 어떤 모양이었을까?
- 회귀분석에서 대부분의 데이터와 떨어져 있는 점을 이상점이라고 한다.

기타 출처 #14

회귀분석에서 이상점의 역할

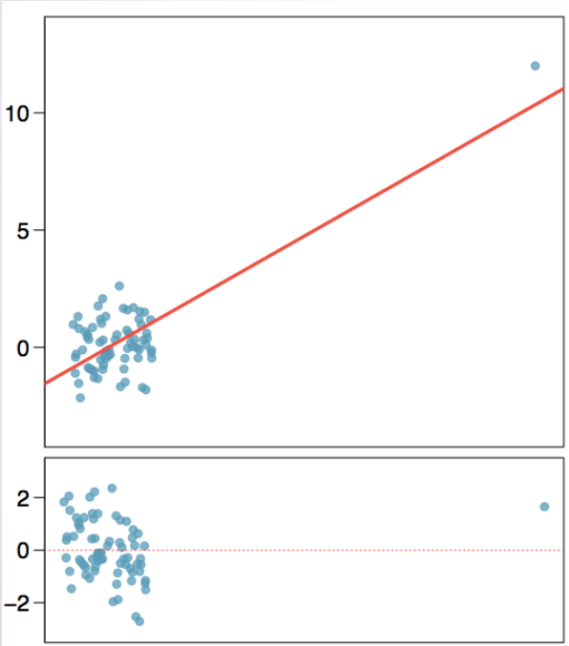


[회귀분석에서의 이상점]
(OpenIntro Statistics, p329)

- 왼쪽 그림에서 이상점은 회귀직선의 기울기에는 영향을 미치지는 않지만 대부분 데이터의 중심에서 수평적으로 떨어져있다.
- 이러한 점을 high leverage point라고 한다.

기타 출처 #15

회귀분석에서 이상점의 역할



[회귀분석에서의 이상점]
(OpenIntro Statistics, p329)

- 그림에서 오른쪽 상단의 한 점이 없었다면 회귀직선은 어떤 모형이었을까?
- 이렇게 회귀직선의 기울기에 영향을 주는 점을 influential point라고 한다.
- 여기서 이점은 high leverage point이기도 하다.

기타 출처 #16

오늘의 강의 요점

- 평균으로의 회귀
- 외삽법
- 이상점

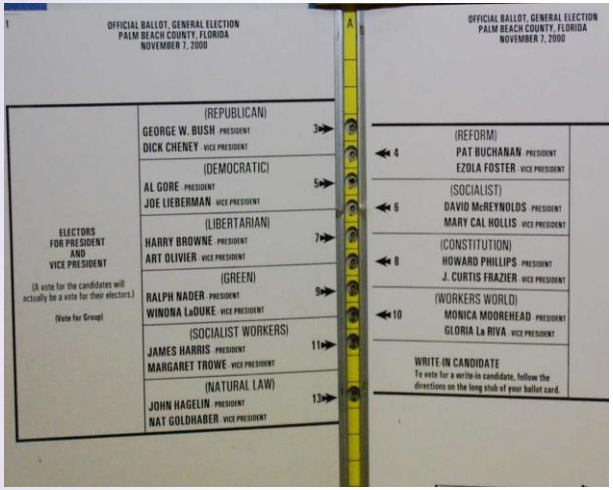
우리 집 가격은 얼마지? – 회귀모형

Lab 6 사례연구: 2000년 미 대선

2000년 미 대선에서는 무슨 일이 있었던 걸까?

- 2000년 미국 대선에서 엘 고어 민주당 후보와 조지 W. 부시 후보간의 치열한 접전이 벌어졌다.
- 막판에 플로리다 주의 투표결과가 박빙으로 이어지면서 부시 후보가 불과 1,784표차로 이기는 것으로 결과가 나오자 재검표에 들어가게 된다.
- 재검표 결과 수작업 재검표에 유효성에 관한 문제로 법정 논쟁이 벌어진 결과 연방대법원이 부시 후보의 손을 들어주어서 조지 W. 부시 후보가 제 44대 미 대통령으로 당선이 확정되었다.

선거 용지가 역사를 바꾼 걸까?



[팜 비치 카운티에서 사용된 나비 모양 투표용지]
(Wikipedia)

- 2000년 미 대선에 고어와 부시 이외의 후보들도 출마하였는데 이 중 뷰캐넌 후보가 팜 비치 카운티에서 유독 득표를 많이 하였다는 사실이 주목을 받았다.
- 그림은 팜 비치 카운티의 2000년 대선에서 사용된 투표용지인데 고어를 지지하는 사람들이 실수로 뷰캐넌 후보에게 투표를 했을 것이라는 주장이 제기되었다.

기타 출처 #17

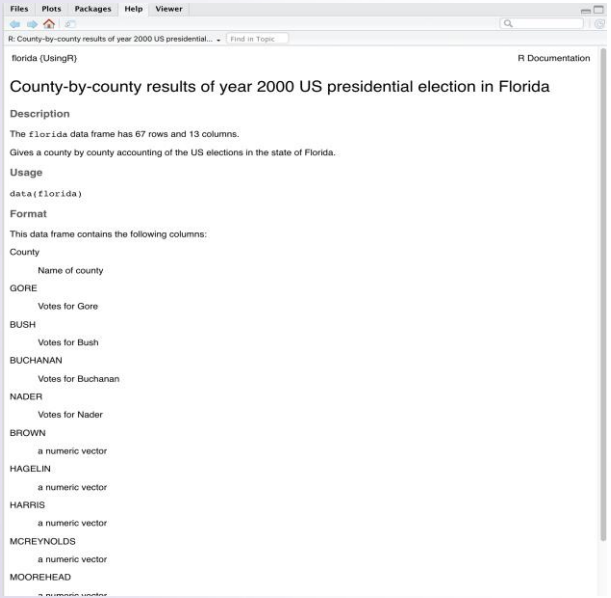
2000 미 대선 플로리다 선거결과를 분석해보자

- Working directory는 한번 지정이 되면 매번 지정할 필요가 없으며 새로운 프로젝트를 시작할 경우 다시 지정하면 된다.
- 먼저 R package “UsingR”을 설치한 후에 command line에 library(UsingR)을 타이핑한다.
- R에서 특정 패키지를 사용하기 위해서는 library(package_name)를 사용하여 패키지를 불러줘야 한다.

2000 미 대선 플로리다 선거결과를 분석해보자

- UsingR 패키지에 내장되어 있는 데이터 셋 “florida”를 사용하기 위해 command line에 attach(florida)를 타이핑 하자.
- 프로그램 종료 시 detach(florida) 사용하여 데이터 셋을 비활성화 하자. 이 부분을 생략할 경우 florida 데이터 셋에 있는 변수와 동일한 변수명을 나중에 사용할 경우 문제가 생길 수 있다.

데이터 셋에 관한 설명은 어디서 찾아야 하나?



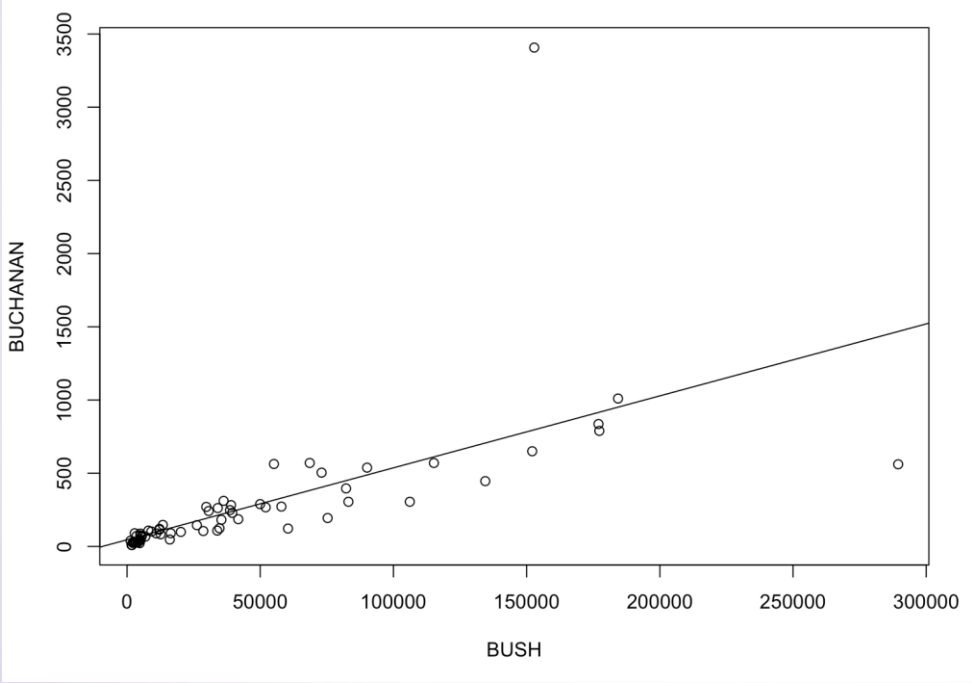
[florida 데이터 셋]

- RStudio의 help 창을 이용하여 florida 데이터 셋에 대한 정보를 얻을 수 있다.
- 플로리다의 67개 카운티 별로 13개의 정보가 기록되어 있다.
- 우리는 이중 BUSH(부시 후보의 득표)을 예측변수로, BUCHANAN(뷰캐넌 후보의 득표)를 반응변수로 사용하여 회귀분석을 실시한다.

2000 미 대선 플로리다 선거결과를 분석해보자

```
> library(UsingR)
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, units


Attaching package: 'UsingR'

The following object is masked from 'package:survival':

    cancer

> attach(florida)
> result.lm <- lm(BUCHANAN ~ BUSH)
> plot(BUSH,BUCHANAN)
> abline(result.lm)
>
```

2000 미 대선 플로리다 선거결과를 분석해보자



[부시 후보의 카운티 별 득표수 vs 뷰캐넌 후보의 카운티 별 득표수]

기타 출처 #18

회귀분석 결과 알아보기

- 아래 output을 보면 회귀직선은 다음과 같이 주어진다.
- $\#\ Buchanan\ vote = 45.29 + 0.0049 \cdot (\#\ Bush\ vote)$

```
> summary(result.lm)

Call:
lm(formula = BUCHANAN ~ BUSH)

Residuals:
    Min      1Q  Median      3Q     Max
-907.50  -46.10  -29.19   12.26 2610.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.529e+01  5.448e+01   0.831    0.409
BUSH        4.917e-03  7.644e-04   6.432 1.73e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353.9 on 65 degrees of freedom
Multiple R-squared:  0.3889,    Adjusted R-squared:  0.3795
F-statistic: 41.37 on 1 and 65 DF,  p-value: 1.727e-08
```

회귀분석 결과 알아보기

```
> summary(result.lm)

Call:
lm(formula = BUCHANAN ~ BUSH)

Residuals:
    Min      1Q  Median      3Q     Max
-907.50  -46.10  -29.19   12.26 2610.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.529e+01  5.448e+01   0.831    0.409
BUSH        4.917e-03  7.644e-04   6.432 1.73e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353.9 on 65 degrees of freedom
Multiple R-squared:  0.3889,    Adjusted R-squared:  0.3795
F-statistic: 41.37 on 1 and 65 DF,  p-value: 1.727e-08
```

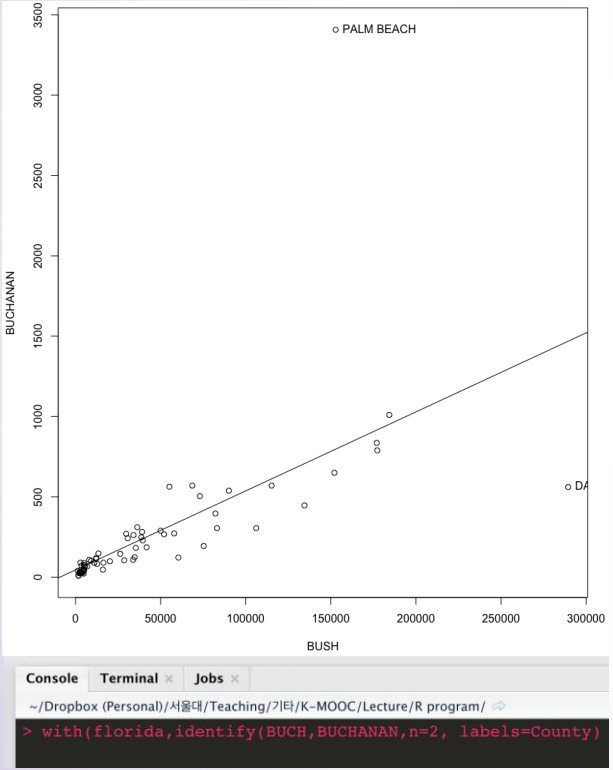
- 여기서 우리가 관심이 있는 것은 기울기가 0인지 여부인다. 뒤의 가설검정에 가서 이 부분을 다시 자세히 알아보자.

회귀분석 결과 알아보기

```
> summary(result.lm)

Call:
lm(formula = BUCHANAN ~ BUSH)

Residuals:
    Min      1Q  Median      3Q     Max
-907.50  -46.10  -29.19   12.26 2610.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.529e+01  5.448e+01   0.831    0.409
BUSH        4.917e-03  7.644e-04   6.432 1.73e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 353.9 on 65 degrees of freedom
Multiple R-squared:  0.3889,    Adjusted R-squared:  0.3795
F-statistic: 41.37 on 1 and 65 DF,  p-value: 1.727e-08
```

- 또 하나의 주목한 요약치는 Multiple R-squared이다. 이 분석에서는 0.3889를 제시하고 있으며 이 의미는 반응변수의 변동(분산)중 39%를 이 회귀모형으로 설명할 수 있다는 것을 의미한다.

이상치 알아보기

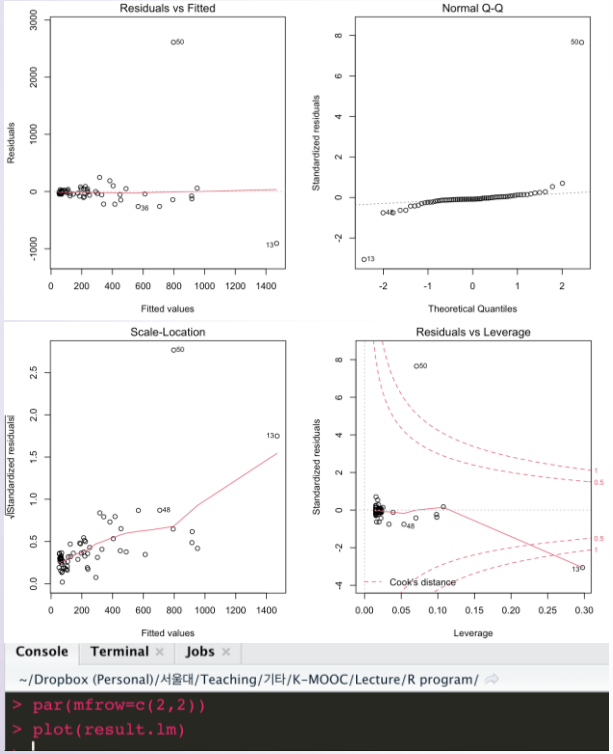


[부시 후보의 카운티 별 득표수 vs 뷰캐넌 후보의 카운티 별 득표수]

- 회귀직선과 유독 떨어진 2개의 점이 어느 카운티에 해당하는지 알고 싶다
- 이 경우 하단의 명령어를 사용하면 된다.
- 이 후에 마우스커서를 해당 점에 이동시킨 후 2개의 점을 **모두** 클릭하면 이름이 나타난다.

기타 출처 #19

회귀모형 진단

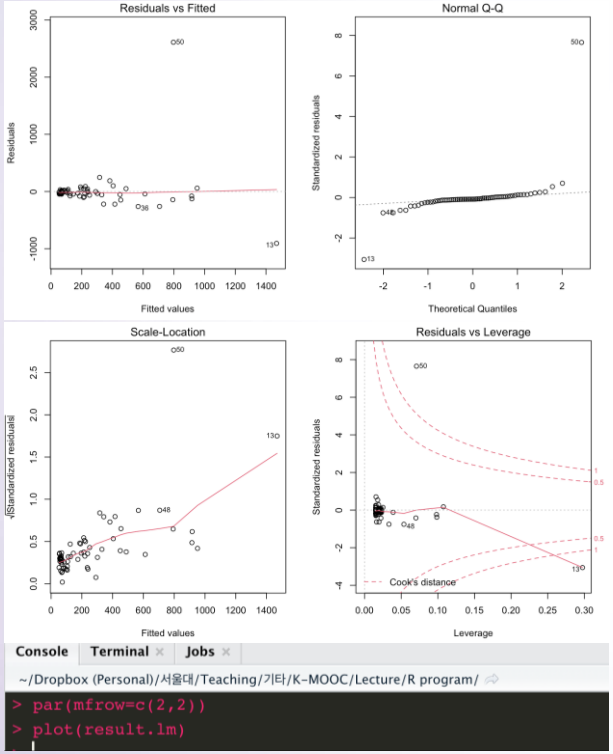


[회귀모형 진단]

기타 출처 #20

- R에서 하단과 같은 간단한 명령어를 통해서 회귀모형 진단을 위해서 4개의 그림을 제공한다.
- 왼쪽 상단의 잔차 vs 반응변수의 추정치의 산점도로 선형관계 가정에 대한 검증을 할 수 있다.

회귀모형 진단



[회귀모형 진단]

기타 출처 #20

- 오른쪽 상단의 경우 정규성 가정을 검증하는 방법으로 오차항이 정규분포를 따른다면 직선모양을 관측할 수 있어야 한다.
- 왼쪽 하단의 그림은 등분산성에 대한 가정을 체크할 수 있다.
- 오른쪽 하단의 그림은 이상점 유무를 탐지하는데 유용하다.

오늘의 강의 요점

- **회귀분석 따라하기**
  - 산점도
  - 기울기와 R^2
  - 회귀모형 진단
  - 이상점 찾기

출처

#1~2 James G. Scott, (2020), Data Science: A Gentle Introduction

#3 D. Spiegelhater, (2019), The Art of Statistics, Penguin Random House

#4 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M, (2019), OpenIntro Statistics, 4th edition, OpenIntro, Inc.

#5~6 James G. Scott, (2018), Data Science: A Gentle Introduction

#7~10 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M, (2019), OpenIntro Statistics, 4th edition, OpenIntro, Inc.

#11~12 Youtube LES https://bit.ly/3683AAf

#13 Wikimedia https://commons.wikimedia.org/wiki/File:Anscombe's_quartet_3.svg#mediaviewer/File:Anscombe's_quartet_3.svg

#14~16 Diez, D.M., Barr, C.D. and Çetinkaya-Rundel, M, (2019), OpenIntro Statistics, 4th edition, OpenIntro, Inc.

#17 Wikipedia https://en.wikipedia.org/wiki/2000_United_States_presidential_election_recount_in_Florida