

Week 8. Simple Regression

Fox., John. 2016. *Applied Regression Analysis and Generalized Linear Models*, 3rd Ed.

Ch. 1. Statistical Models and Social Science

Ch. 2. What Is Regression Analysis?

Ch. 5. Linear Least-Squares Regression, pp. 82-91.

Sanghoon Park

Department of Political Science
University of South Carolina

2020-10-25

Introduction

Y : 측정된 몸무게 / X : 보고된 몸무게

- 두 변수의 관계를 보여주는 선: $Y = A + BX$
- 아무리 두 변수의 관계가 강력한 선형 관계를 보일지라도 각 관측치들이 모두 선 위에 완벽하게 놓이는 선은 존재할 수 없음.
- 왜? 현실세계에는 우리가 관측하지 못한 요인들이 두 변수의 관계에 영향을 미칠 수 있기 때문

불확실성: 잔차(residual)

- 실제 관측한 데이터들의 관계에는 불확실성이 개입될 수밖에 없음.
- $n = 101$ 개인 표본에서 i 번째 관측치에 대한 회귀식:
$$Y_i = A + BX_i + E_i = \hat{Y}_i + E_i$$
- 이때, $\hat{Y}_i = A + BX_i$ 는 각 관측치 i 에 대한 예측값(fitted value)

Introduction

Y : 측정된 몸무게 / X : 보고된 몸무게

- 두 변수의 관계를 보여주는 선: $Y = A + BX$
- 아무리 두 변수의 관계가 강력한 선형 관계를 보일지라도 각 관측치들이 모두 선 위에 완벽하게 놓이는 선은 존재할 수 없음.
- 왜? 현실세계에는 우리가 관측하지 못한 요인들이 두 변수의 관계에 영향을 미칠 수 있기 때문

불확실성: 잔차(residual)

- 실제 관측한 데이터들의 관계에는 불확실성이 개입될 수밖에 없음.
- $n = 101$ 개인 표본에서 i 번째 관측치에 대한 회귀식:
$$Y_i = A + BX_i + E_i = \hat{Y}_i + E_i$$
- 이때, $\hat{Y}_i = A + BX_i$ 는 각 관측치 i 에 대한 예측값(fitted value)

Introduction

Y : 측정된 몸무게 / X : 보고된 몸무게

- 두 변수의 관계를 보여주는 선: $Y = A + BX$
- 아무리 두 변수의 관계가 강력한 선형 관계를 보일지라도 각 관측치들이 모두 선 위에 완벽하게 놓이는 선은 존재할 수 없음.
- 왜? 현실세계에는 우리가 관측하지 못한 요인들이 두 변수의 관계에 영향을 미칠 수 있기 때문

불확실성: 잔차(residual)

- 실제 관측한 데이터들의 관계에는 불확실성이 개입될 수밖에 없음.
- $n = 101$ 개인 표본에서 i 번째 관측치에 대한 회귀식:
$$Y_i = A + BX_i + E_i = \hat{Y}_i + E_i$$
- 이때, $\hat{Y}_i = A + BX_i$ 는 각 관측치 i 에 대한 예측값(fitted value)

Introduction

Fox (2016, 84)는 단순순회귀모델의 개념을 시각화를 통해 보여줌.

- $E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$
- 잔차(E_i)란 실제 개별 관측치(Y_i 로부터 모델로 예측한 값(\hat{Y}_i)의 차이

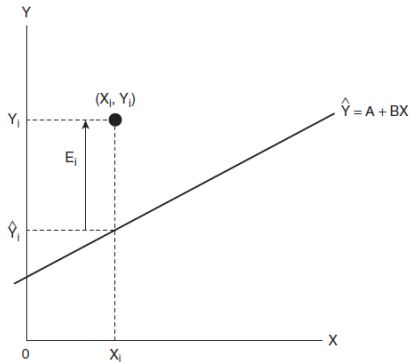


Figure 1: i 번째 관측치에 대한 잔차 E_i 를 보여주는 Y 에 대한 X 의 회귀곡선

Residuals

데이터에 잘 들어맞는 회귀선이란 잔차의 크기가 작은 회귀선

- 하지만 이때 '작다'는 것은 단 하나의 값에 대한 개별 잔차가 작다는 것을 의미하는 것이 아님.
- 잔차의 절대값, 즉 개별 관측치와 모델의 예측값 간의 거리가 모든 관측치에 대해 평균적으로 작은 경우를 의미.
- 이를 위해서 우리는 잔차의 총합(sum of residuals; $\sum_{i=1}^n E_i$)의 값이 작을 때, 모델이 정확한 것으로 기대

만약 변수들의 평균(\bar{X} , \bar{Y})을 지나는 선이 존재한다고 할 때,

- 평균을 지나는 선은 다음과 같이 나타낼 수 있음: $\bar{Y} = A + B\bar{X}$ 의 의미.
- 회귀곡선과 마찬가지로 평균을 지나는 선과 개별 관측치와의 관계도 수식으로 나타낼 수 있음.
- $Y_i - \bar{Y} = B(X_i - \bar{X}) + E_i$

Least-Squares Fit

모든 관측치와 변수들의 평균을 지나는 선 간의 잔차를 총합 역시 간단하게 구할 수 있음.

- $\sum_{i=1}^n E_i = \sum (Y_i - \bar{Y}) - B \sum (X_i - \bar{X}) = 0 - B \times 0 = 0$

즉, 이 결과는 특정 모델(회귀 or 평균)으로 예측한 값과 개별 관측치들 간 잔차의 합은 $0(\sum E_i = 0)$ 이라는 것을 보여줌.

- 왜냐하면 평균을 지나는 선 아래에 위치한 값, 위로 위치한 값들의 잔차는 결국 \pm 로 상쇄되어 0으로 수렴
- 문제는 서로 다른 부호를 가진 잔차들이 상쇄되어 0이 되어버린다고 할 때, 서로 다른 모델 중에 어떤 것이 더 '작은 잔차'를 가지고 더 '정확한 모델'인지 비교하기 어려움.

그렇다면 '부호'를 없애주는 방법을 생각해볼 수 있음.

1. 잔차의 절대값을 이용하는 방법(least-absolute-value (LAV) regression)

- 잔차의 절대값의 합($\sum |E_i|$)의 합을 최소화시켜주는 A 와 B 를 찾는 것

2. 잔차의 제곱값을 이용하는 방법(least-squares-criterion)

- 잔차의 제곱값의 합($\sum E_i^2$)의 합을 최소화시켜주는 A 와 B 를 찾는 것

Least-Squares Fit

모든 관측치에 대한 잔차의 제곱합을 최소화해주는 A와 B의 값을 구하고자 함.

- $S(A, B) = \sum_{i=1}^n E_i^2 = \sum (Y_i - A - BX_i)^2$
- 주어진 데이터 $\{X_i, Y_i\}, i = 1, \dots, n$ 에 대해 특정한 잔차의 제곱합($\sum E_i^2$)에 대응하는 A와 B의 가능한 값을 나타내는 함수라고 할 수 있음.
- 우리는 잔차의 제곱합이 최소가 되는 A와 B의 짝을 찾고자 하는 것 (?).

가장 직접적으로 최소제곱합 접근법에 따른 계수값을 구하는 것은 위의 제곱합 함수 (sum-of-squares function)에 편미분을 취하는 것

- $\frac{\partial S(A, B)}{\partial A} = \sum (-1)(2)(Y_i - A - BX_i)$
- $\frac{\partial S(A, B)}{\partial B} = \sum (-X_i)(2)(Y_i - A - BX_i)$
- 간단하게 말하면 기울기 값을 구하는 것이라고 생각할 수 있음.
- Fox (2016, 86)의 측정된 몸무게(Y)와 보고된 몸무게(X)에 관한 예제를 살펴보자.

Least-Squares Fit

$$n = 101$$

$$\bar{Y} = \frac{5780}{101} = 57.228$$

$$\bar{X} = \frac{5731}{101} = 56.743$$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 4435.9$$

$$\sum (X_i - \bar{X})^2 = 4539.3$$

$$B = \frac{4435.9}{4539.3} = 0.97722$$

$$A = 57.228 - 0.97722 \times 56.743 = 1.7776$$

Least-Squares Fit

이 식이 의미하는 것은 $\widehat{\text{측정 몸무게}} = 1.78 + 0.977 \times \text{보고 몸무게}$

- $B = 0.977$: 보고된 몸무게 1kg 증가는 평균적으로 측정된 몸무게 약 0.977 kg 증가와 관련이 있다는 의미.
- $A = 1.78$: X , 즉 보고된 몸무게가 0일 때의 측정된 몸무게의 값을 말하는데, 현실적으로 몸무게가 0인 사람이 존재할 수 없으므로 유의미하게 해석하는 값은 아님.
- 만약 보고된 몸무게가 실제 측정된 몸무게에 대해 편향되지 않은 (unbiased) 예측값이었다면, 우리는 $\hat{Y} = X$ 라는 식을 얻었을 것이고, 절편값은 0에 기울기 값은 1이었을 것.

Simple Correlation

잔차의 제곱합을 최소화하는 선을 그렸다고 할 때, 이제 우리의 관심사는 얼마나 그 선이 실제 관측치가 퍼져있는 것에 들어맞는지를 확인하는 것

- 잔차의 표준편차(standard deviation of the residuals); 회귀의 표준오차(standard error of the regression)¹

- 잔차의 분산은 $n - 2$ 의 자유도로 정의되므로 다음과 같이 나타낼 수 있음:

$$S_E^2 = \frac{\sum E_i^2}{n-2}.$$

- 따라서 잔차의 표준편차는 이 분산의 제곱근 값이 됨: $S_E = \sqrt{\frac{\sum E_i^2}{n-2}}.$

잔차의 표준편차는 실제 관측치를 잔차의 제곱합으로 계산한 회귀곡선으로 예측을 했을 때, 우리가 '평균적으로' 얻을 수 있는 오차를 의미함.

- 이러한 잔차는 대략적으로 정규분포를 따름.

¹Fox (2016, 87)의 각주 11에서도 밝히고 있듯이, 보통 표준오차는 통계치들의 표집분포에서 추정된 표준편차(e.g., 개별 표본들의 평균이 가지는 분포의 표준편차)를 의미함. 하지만 잔차의 표준편차를 잔차의 표준오차와 서로 교환가능한 개념으로 만연하게 쓰기도 함. 엄밀히 말하면 잔차의 표준오차라는 표현은 잘못된 개념임.

Simple Correlation

회귀의 표준오차(=잔차의 표준편차)와 달리 상관계수(correlation coefficient)는 상대적인 적합도(fit)를 보여줌.

- X 와 Y 에 대한 선형관계로 예측을 했을 때, X 없이 예측했을 때에 비해 얼마나 Y 에 대한 예측이 개선되었는가를 의미
- 상대적인 적합도라는 것은 기준점(baseline)이 필요하다는 것을 의미.
- 따라서 우리는 X 없이 Y 를 어떻게 예측할 수 있는지 생각해보아야 함.

Fox (2016, 88-89)은 X 없이 Y 를 예측할 때, 가장 효율적인 방법— Y 의 평균으로 예측하는 것을 증명하는 과정을 보여주고 있음.

- 앞서 평균이 중요한 이유는 추가적인 정보가 없을 때, 특정 집단을 대표하는 값으로 사용할 수 있기 때문이라고 한 바 있음.
- 따라서 만약 특정한 X 가 없다면 우리는 다른 값을 무턱대고 찍기보다는 \bar{Y} 로 Y 를 예측하고자 할 것임.
- 그렇다면 $\sum(Y_i - \hat{Y}_i)^2 \leq \sum(Y_i - \bar{Y})^2$ 라면 우리는 X 를 가지고 예측했을 때 잔차 제곱합이 더 작으므로 평균보다는 적어도 X 를 가지고 Y 를 추정하는 것이 더 효율적이라고 할 수 있음.

Simple Correlation

$$\sum E_i'^2 = \sum (Y_i - \bar{Y})^2 = \text{평균으로 예측 시 잔차 제곱합}$$

=Total sum of squares; TSS

$$\sum E_i^2 = \sum (Y_i - \hat{Y})^2 = \text{설명변수로 예측 시 잔차 제곱합}$$

=residual sum of squares; RSS

이때, TSS 와 RSS 의 차이를 회귀 제곱합(regression sum of squares; RegSS)라고 하면($RegSS \equiv TSS - RSS$)은 선형회귀로 인해 감소한 오차의 제곱(squared error)의 크기를 보여줌.

- 따라서 TSS 에 대한 $RegSS$ 의 비율은 평균으로 예측했을 때의 잔차에 비해 회귀분석으로 구한 잔차의 값이 줄어든 크기를 평균 예측 잔차와 비교한 것으로써, 오차의 제곱의 비율적 감소(the proportional reduction in squared error)를 보여줌.
- 그리고 그 값을 우리는 상관계수의 제곱값으로 나타냄: $r^2 \equiv \frac{RegSS}{TSS}$.
- 회귀계수 B 가 +면 r^2 의 $+\sqrt{\quad}$ 또는 -면 $-\sqrt{\quad}$

Simple Correlation

상관계수의 함의

- 만약 Y 와 X 사이에 완벽한 양의 선형 관계가 존재(잔차가 0, $B > 0$), $r = 1$.
- 만약 Y 와 X 사이에 완벽한 음의 선형 관계가 존재(잔차가 0, $B < 0$), $r = -1$.
- 만약 Y 와 X 사이에 선형 관계가 없다면($RSS = TSS$, $RegSS = 0$), $r = 0$.
- $-1 \leq r \leq 1$ 사이에서 r 은 두 변수 간의 선형관계의 방향성을 알려줌.
- r^2 는 X 로 수행한 선형회귀분석이 설명하는 Y 의 총 변동량의 비율이라고 할 수 있음: X 로 Y 를 몇 % 설명할 수 있는가?

선형회귀모델은 종속변수의 총 변동량을 “설명된” 부분과 “설명되지 않은” 부분으로 나누어 살펴볼 수 있음.

- King, Keohane, and Verba (1994) 식 표현대로라면 체계적(systematic) 요인과 비체계적 요인
- 혹은 예측값(fitted values)과 잔차(residuals)
 - 이러한 접근법을 회귀분석에 대한 분산분석(analysis of variance)라고 함.

Simple Correlation

두 확률변수 X 와 Y 의 상관계수는 수식으로 $\rho = \sigma_{XY} / \sigma_X \sigma_Y$ 로 나타낼 수도 있음.

- σ_X : X 의 표준편차
- σ_Y : Y 의 표준편차

개념적으로 말하자면 상관계수란 두 변수가 각각 분포해 있는 정도에 비해 두 분포가 공유하고 있는 분포 정도(공분산)의 비율을 구하는 것

- 먼저 표본의 공분산(sample covariance)를 계산

$$S_{XY} \equiv \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- 이때 우리는 상관계수 r 을 다음과 같이 계산할 수 있음.

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Simple Correlation

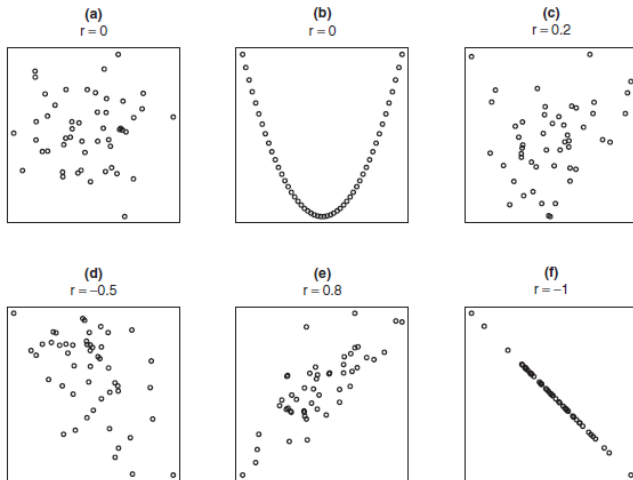


Figure 2: 상관계수의 값에 따른 산포도: (a)와 (b)는 $r = 0$, (c)는 $r = 0.2$, (d)는 $r = -0.5$, (e)는 $r = 0.8$, (f)는 $r = -1$ 일 경우를 보여줌. (b)를 제외한 나머지 모든 패널의 $n = 50$.

Basic Assumptions

단순선형회귀모델(SLR)이 편향되지 않은(unbiased) 결과를 추정하기 위한 조건

- 다중선형회귀모델(MLR)에서 이 가정들은 확장될 것임.

전통적인 선형회귀분석의 가정들은 교차사례(cross-sectional) 자료에 적용가능

- 시간에 따른 변화; 시계열(time-series)이 포함되면 주요 가정이 위배될 가능성이 존재 → 추가적인 고려 필요

Basic Assumptions

주요 가정

- SLR1. 모집단에서 모수(parameters)의 관계가 선형(linear)
- SLR2. 무작위로 추출된 표본(randomized sample)
- SLR3. 설명변수(X)와 오차항(잔차; u)은 서로 독립
 - $E(X|u) = 0$
 - 이 가정이 성립되지 않을 시, 우리는 모델 특징에 문제가 있다고 생각할 수 있음.
 - 예를 들어, 종속변수에 영향을 미치는 적절한 변수를 모델에 포함시키지 않았을 가능성 (누락변수의 문제; OVB)²
- SLR4. 오차항의 동분산성(homoskedasticity)
 - 주어진 X 에 대해 오차항의 조건분포가 일정해야 함.
 - 단, 오차항에 이분산성(heteroskedasticity)이 있더라도 반드시 편향된 결과라는 것은 아님.
 - 다만 추정치의 효율성(efficiency)가 낮다는 의미.

²Omitted Variable Bias Problem

Example

가상의 데이터: Quality of Governance

- Y : 2017년 국가들의 경제성장률
- X : 2017년 국가들의 민주주의 수준

표본의 기술통계

	n	mean	sd
ccodecow	148	459.905	228.985
cname*	148	74.500	42.868
p_polity2	148	4.351	6.016
wdi_gdpagr	148	11.290	11.386

데이터의 모습

ccodecow	p_polity2	wdi_gdpagr
700	-1	20.467
339	9	19.014
615	2	12.272
540	-2	10.017

Example

양변인관계 (bivariate relationship)의 탐색

- 두 변수 모두 연속형이므로, 산포도 (scatter plot)으로 두 변수의 관계를 대략 살펴볼 수 있음.

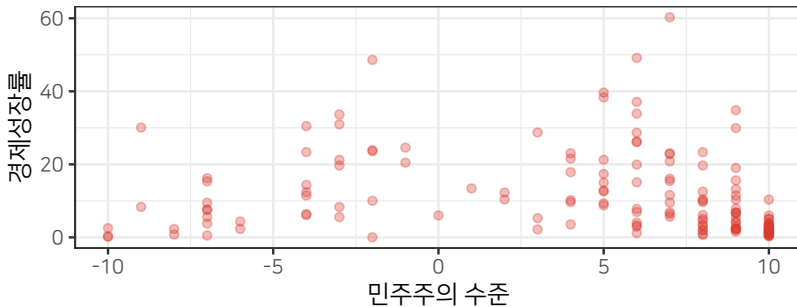


Figure 3: Bivariate relationship between democracy and economic growth in 2017

Example

단순회귀선형모델의 적용

- 개념적으로 우리가 구성하는 모델은 다음과 같음.

$$\text{경제성장률}_i = \beta_0 + \beta_1(\text{민주주의 수준}_i) + \epsilon_i$$

위의 모델을 R로 재구성할 수 있음.

```
model <- lm(wdi_gdpagr ~ p_polity2,  
            data = QOG_summary)
```

Example

모델 분석 결과는 `summary()` 함수로 확인

```
summary(model)
```

하지만 `summary()`은 콘솔창에 그 결과를 텍스트로 보여주기 때문에 추가적인 분석에 어려움이 있음.

- 모델 결과를 데이터로 저장할 수 있음: `broom()` 사용

```
tidy_model <- broom::tidy(model)
head(tidy_model) %>% knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.51	1.148	10.901	0.000
p_polity2	-0.28	0.155	-1.809	0.072

Example

tidymodels 패키지의 `parsnip()`을 이용하여 선형회귀분석을 수행할 수 있음.

```
library(parsnip)
parsnip_model <- linear_reg() %>%
  set_engine("lm") %>%
  fit(wdi_gdpagr ~ p_polity2,
      data = QOG_summary)
```

`parsnip()`으로 분석한 결과 역시도 `broom` 패키지의 `tidy()` 함수를 통해 데이터화할 수 있음.

```
broom::tidy(parsnip_model) %>%
  head(n = 3) %>%
  knitr::kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	12.51	1.148	10.901	0.000
p_polity2	-0.28	0.155	-1.809	0.072

Example

분석결과의 간략한 이해

- `estimate`가 추정치, `std.error`가 표집에 따라 그 추정치가 가지는 불확실성을 보여줌.
- `statistic`은 `t-statistics`를 의미하며, 1.96 이상일 때, `p-value`는 0.05 이하로 나타남 (전통적인 통계적 유의수준).
- 민주주의는 경제성장률과 통계적으로 유의미한 선형관계를 가지고 있지 않다는 영가설을 기각할만한 충분한 근거가 없음.

분석 결과를 시각화를 통해 보여줄 수 있음.

- 먼저 신뢰구간을 계산함.
 - 직접 추정치에 $\pm 1.96 \times \text{Standard Errors}$ 로 구할 수 있음.
 - `broom()`의 `confint_tidy()` 함수로 간단히 구할 수 있음.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	12.51	1.148	10.901	0.000	10.242	14.778
p_polity2	-0.28	0.155	-1.809	0.072	-0.586	0.026

Example

선형회귀모델 결과의 시각화

```
tidy_model %>%  
  ggplot(aes(x = term, y = estimate)) +  
  geom_pointrange(aes(ymin = conf.low,  
                      ymax = conf.high),  
                 size = 0.5) + labs(x = "") +  
  geom_hline(yintercept = 0, color = "red")
```

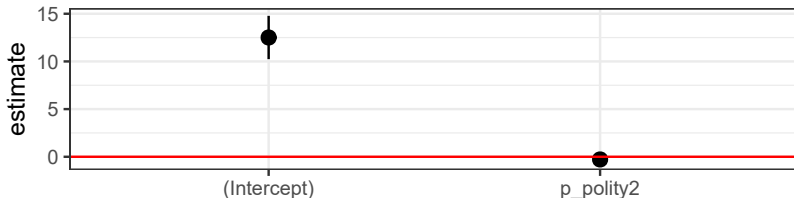


Figure 4: Simple linear regression model of democracy and economic growth in 2017

- Fox, John Jr. 2016. Applied Regression Analysis and Generalized Linear Models. 3 ed. CA: Thousand Oaks, SAGE Publications.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton, N.J.: Princeton University Press.

Table of Contents

Introduction

Residuals

Least-Squares Fit

Simple Correlation

Simple Regression