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

2020-10-24

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 i번째 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 i에 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 i번째 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 i에 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 i번째 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 i에 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_i로부터 모델로 예측한 값($\hat{Y}_i$)의 차이

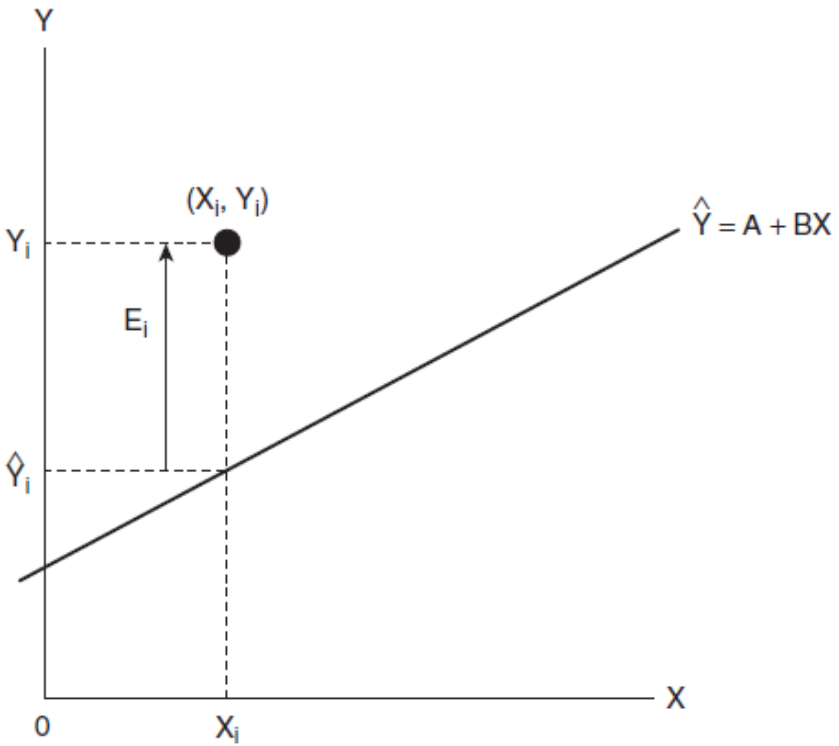

Figure 1: i번째 관측치에 대한 잔차 E_i를 보여주는 Y에 대한 X의 회귀곡선

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 A와 B를 찾는 것

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

- 잔차의 제곱값의 합($\sum E_i^2$)의 합을 최소화시켜주는 A와 B를 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 $\{X_i, Y_i\}, i = 1, \ldots, n$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 funcction)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 $\widehat{\text{측정 몸무게}} = 1.78 + 0.977 \times$ 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 최소로하는 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 씨, 우리가 '평균적으로' 얻을 수 있는 오차를 의미함.

- 이러한 잔차는 대략적으로 정규분포를 따름.

[1] Fox (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.

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

Fox (2016, 88-89)은 X 없이 Y를 예측할 때, 가장 효율적인 방법—Y의 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 X가 없다면 우리는 다른 값을 무턱대고 찍기보다는 $\bar{Y}$로 Y를 예측하고자 할 것임.
- 그렇다면 $\sum(Y_i - \hat{Y}_i)^2 \leq \sum(Y_i - \bar{Y})^2$라면 우리는 X를 가지고 예측했을 때 잔차 제곱합이 더 작으므로 평균보다는 적어도 X를 가지고 Y를 추정하는 것이 더 효율적이라고 할 수 있음.

Simple Correlation

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

$$= \text{Total sum of squares; TSS}$$

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

$$= \text{residual sum of squares; RSS}$$

이때, TSS와 RSS의 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 TSS에 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: $r^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$.
- 회귀계수 B가 +면 r^2의 $+\sqrt{\ }$ 또는 −면 $-\sqrt{\ }$

Simple Correlation

상관계수의 함의

- 만약 Y와 X 사이에 완벽한 양의 선형 관계가 존재(잔차가 0, $B > 0$), $r = 1$.
- 만약 Y와 X 사이에 완벽한 음의 선형 관계가 존재(잔차가 0, $B < 0$), $r = -1$.
- 만약 Y와 X 사이에 선형 관계가 없다면($RSS = TSS$, $RegSS = 0$), $r = 0$.
- $-1 \leq r \leq 1$ 사이에서 r은 두 변수 간의 선형관계의 방향성을 알려줌.
- r^2는 X로 수행한 선형회귀분석이 설명하는 Y의 총 변동량의 비율이라고 할 수 있음: X로 Y를 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 X와 Y의 상관계수는 수식으로 $\rho = \sigma_{XY}/\sigma_X\sigma_Y$로 나타낼 수도 있음.

- σ_X: X의 표준편차
- σ_Y: Y의 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 r을 다음과 같이 계산할 수 있음.

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

Simple Correlation

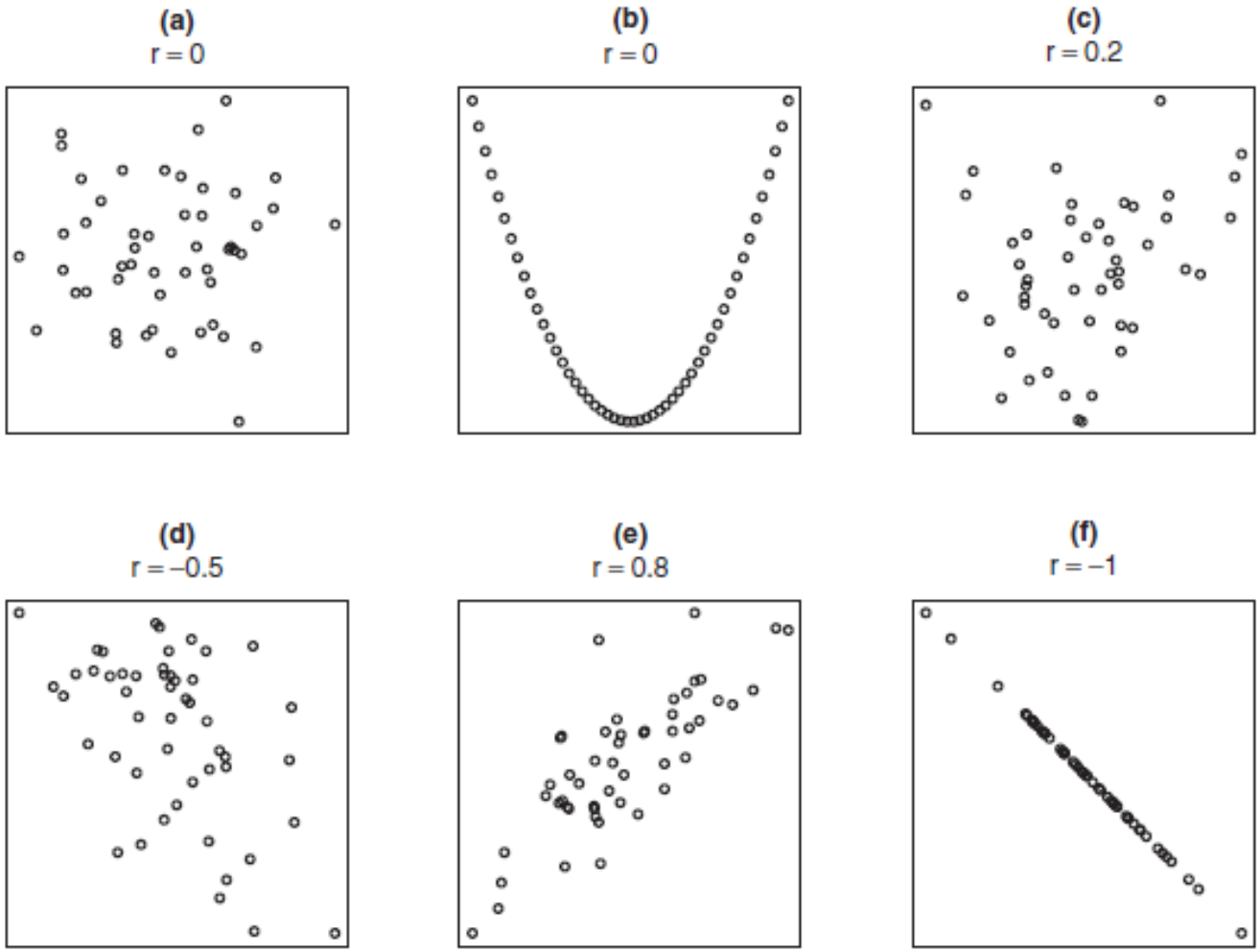

Figure 2: 상관계수의 값에 따른 산포도: (a)와 (b)는 $r = 0$, (c)는 $r = 0.2$, (d)는 $r = -0.5$, (e)는 $r = 0.8$, (f)는 $r = -1$일 경우를 보여줌. (b)를 제외한 나머지 모든 패널의 $n = 50$.

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