Week 10. Multivariate Linear Regression I. Theory

Fox., John. 2016. *Applied Regression Analysis and Generalized Linear Models*, 3rd Ed.
Ch. 5. Linear Least-Squares Regression, pp. 92-102.
Fox., John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*, 3rd Ed.
Ch. 4. Fitting Linear Models

Sanghoon Park

Department of Political Science
University of South Carolina

2020-11-15

Introduction

다중선형회귀모델(Multivariate Linear Regression Models); 이하 MLR

- 단순선형회귀모델(Simple Linear Regression Models, SLR)은 종속변수 하나와 예측변수 하나 간의 대응을 나타냄
  - $Y = \beta_0 + \beta_1 X + \epsilon$
- 반면, MLR은 종속변수와 하나 이상의 예측변수 간의 대응관계를 나타냄
  - $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$

먼저 예측변수가 두 개인 MLR에 대해서 살펴보면 다음과 같음

- $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- 종속변수의 예측값, $\hat{Y}$ 는 우변의 함수로 결정된다는 의미

Introduction

하지만 현실 세계에서 우리의 예측값이 완벽하게 종속변수를 예측하리라고는 기대하기 어려움

- 따라서 실제 존재하는 데이터의 관측값, Y_i와 모델에 의한 예측값, $\hat{Y}_i$ 사이에는 차이가 존재할 것 $\rightarrow$ 잔차(residuals)
- $\epsilon_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$

SLR과 마찬가지로 MLR에서 우리는 이 잔차의 제곱합이 최소가 되는 $\beta_0, \beta_1, \beta_2$를 구하는 것이 목표

- $S(\beta_0, \beta_1, \beta_2) = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$
- SLR에서처럼 잔차의 제곱합에 대해 편미분을 통해 각각의 회귀계수값을 구할 수 있음
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_0} = \sum (-1)(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} = \sum (-X_{i1})(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_2} = \sum (-X_{i2})(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$

Introduction

하지만 현실 세계에서 우리의 예측값이 완벽하게 종속변수를 예측하리라고는 기대하기 어려움

- 따라서 실제 존재하는 데이터의 관측값, Y_i와 모델에 의한 예측값, $\hat{Y}_i$ 사이에는 차이가 존재할 것 $\rightarrow$ 잔차(residuals)
- $\epsilon_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$

SLR과 마찬가지로 MLR에서 우리는 이 잔차의 제곱합이 최소가 되는 $\beta_0, \beta_1, \beta_2$를 구하는 것이 목표

- $S(\beta_0, \beta_1, \beta_2) = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$
- SLR에서처럼 잔차의 제곱합에 대해 편미분을 통해 각각의 회귀계수값을 구할 수 있음
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_0} = \sum (-1)(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} = \sum (-X_{i1})(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_2} = \sum (-X_{i2})(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$

Introduction

하지만 현실 세계에서 우리의 예측값이 완벽하게 종속변수를 예측하리라고는 기대하기 어려움

- 따라서 실제 존재하는 데이터의 관측값, Y_i와 모델에 의한 예측값, $\hat{Y}_i$ 사이에는 차이가 존재할 것 $\rightarrow$ 잔차(residuals)
- $\epsilon_i = Y_i - \hat{Y}_i = Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})$

SLR과 마찬가지로 MLR에서 우리는 이 잔차의 제곱합이 최소가 되는 $\beta_0, \beta_1, \beta_2$를 구하는 것이 목표

- $S(\beta_0, \beta_1, \beta_2) = \sum \epsilon_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$
- SLR에서처럼 잔차의 제곱합에 대해 편미분을 통해 각각의 회귀계수값을 구할 수 있음
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_0} = \sum (-1)(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_1} = \sum (-X_{i1})(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$
  - $\frac{\partial S(\beta_0, \beta_1, \beta_2)}{\partial \beta_2} = \sum (-X_{i2})(2)(Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})$

Introduction

이때, 잔차의 제곱합을 최소로 하는 회귀계수, β_1, β_2 는 어디까지나 다음의 조건이 성립할 경우에 단일하게 정의됨

$$\sum X_1^{*2} \sum X_2^{*2} \neq (\sum X_1^* X_2^*)^2$$

- 이 조건의 의미는 "X_1 과 X_2 가 서로 완벽한 상관관계에 있지 않거나" 또는 "설명변수가 모두 변하는 값을 가지는 경우"에만 우리는 잔차의 제곱합을 최소화하는 회귀계수값을 가진다는 것을 의미
- 만약 두 예측변수가 완벽하게 상관관계에 있다면? **공선성(colinear)**이 존재한다고 할 수 있음.
  - 간단하게 생각해보면 X_1 과 X_2 가 완벽하게 상관관계, 즉 서로가 서로를 완벽하게 예측할 수 있는 변수라면 둘 중 하나만 써도 무방함
  - 혹은 높은 수준의 상관관계를 가질 경우, 각 예측변수가 다른 예측변수와 관계없이 독립적으로 종속변수에 미치는 효과가 작다는 것을 의미 $\rightarrow$ 과연 변수로 투입하는 의미가 있을까?

Introduction

이때, 잔차의 제곱합을 최소로 하는 회귀계수, β_1, β_2 는 어디까지나 다음의 조건이 성립할 경우에 단일하게 정의됨

$$\sum X_1^{*2} \sum X_2^{*2} \neq (\sum X_1^* X_2^*)^2$$

- 이 조건의 의미는 "X_1 과 X_2 가 서로 완벽한 상관관계에 있지 않거나" 또는 "설명변수가 모두 변하는 값을 가지는 경우"에만 우리는 잔차의 제곱합을 최소화하는 회귀계수값을 가진다는 것을 의미
- 만약 두 예측변수가 완벽하게 상관관계에 있다면? **공선성(colinear)**이 존재한다고 할 수 있음.
  - 간단하게 생각해보면 X_1 과 X_2 가 완벽하게 상관관계, 즉 서로가 서로를 완벽하게 예측할 수 있는 변수라면 둘 중 하나만 써도 무방함
  - 혹은 높은 수준의 상관관계를 가질 경우, 각 예측변수가 다른 예측변수와 관계없이 독립적으로 종속변수에 미치는 효과가 작다는 것을 의미 → 과연 변수로 투입하는 의미가 있을까?

Introduction

이때, 잔차의 제곱합을 최소로 하는 회귀계수, β_1, β_2 는 어디까지나 다음의 조건이 성립할 경우에 단일하게 정의됨

$$\sum X_1^{*2} \sum X_2^{*2} \neq (\sum X_1^* X_2^*)^2$$

- 이 조건의 의미는 "X_1 과 X_2 가 서로 완벽한 상관관계에 있지 않거나" 또는 "설명변수가 모두 변하는 값을 가지는 경우"에만 우리는 잔차의 제곱합을 최소화하는 회귀계수값을 가진다는 것을 의미
- 만약 두 예측변수가 완벽하게 상관관계에 있다면? **공선성(colinear)**이 존재한다고 할 수 있음.
    - 간단하게 생각해보면 X_1 과 X_2 가 완벽하게 상관관계, 즉 서로가 서로를 완벽하게 예측할 수 있는 변수라면 둘 중 하나만 써도 무방함
    - 혹은 높은 수준의 상관관계를 가질 경우, 각 예측변수가 다른 예측변수와 관계없이 독립적으로 종속변수에 미치는 효과가 작다는 것을 의미 $\rightarrow$ 과연 변수로 투입하는 의미가 있을까?

Introduction

Fox (2016, 94), Duncan's occupational prestige data

$$
\begin{aligned}
n &= 45 \\
\bar{Y} &= \frac{2146}{45} = 47.689 \\
\bar{X_1} &= \frac{2365}{45} = 52.556 \\
\bar{X_2} &= \frac{1884}{45} = 41.867 \\
\sum X_1^{*2} &= 38,971 \\
\sum X_2^{*2} &= 26,271 \\
\sum X_1^* X_2^* &= 23,182 \\
\sum X_1^* Y^* &= 35,152 \\
\sum X_2^* Y^* &= 28,383
\end{aligned}
$$

- $\widehat{\text{Prestige}} = -6.065 + 0.5458 \times \text{Education} + 0.5987 \times \text{Income}$

MLR vs. SLR

SLR과 MLR이 잔차의 제곱합을 최소화하는 회귀계수값을 구한다는 점에서는 비슷함

- 하지만 해석에 있어서는 중요한 차이가 존재
- MLR에서 설명변수의 계수값: 부분 계수 (partial coefficients)
- SLR에서 설명변수의 계수값: 한계 효과 (marginal effects)를 보여줌
    - 설명변수의 한 단위 변화가 종속변수에 미치는 영향을 보여줌

SLR과 달리 MLR의 각 변수의 계수값은 "다른 예측변수의 값을 고정했을 때 (holding constant)의 예측변수 값의 한 단위 증가가 종속변수에 미치는 효과를 보여줌.

- 여러 개의 예측변수들 간에는 불가피하게 통계적으로 서로 "겹치는 부분이 존재할 수 있음"
- 따라서 MLR은 예측변수 하나가 종속변수에 미치는 효과가 아니라 다른 예측변수들의 영향력을 제외하여 순수하게 해당 예측변수만의 변량(variations)이 종속변수에 미치는 부분적 관계를 살펴보는 것

MLR coefficients

MLR: $\hat{Y} + \beta_0 + \beta_1 X_1 + \beta_2 X_2$ 가 있다고 할 때, 편미분을 통해 MLR의 각 예측변수의 부분 회귀계수를 구할 수 있음

- X_2 가 변하지 않는다고 할 때, X_1 의 Y 에 대한 부분적 효과: $\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$
- X_1 이 변하지 않는다고 할 때, X_2 의 Y 에 대한 부분적 효과: $\frac{\partial \hat{Y}}{\partial X_2} = \beta_2$

따라서 위의 Duncan의 자료를 이용한 회귀분석 결과를 해석해보면 다음과 같이 나타낼 수 있음

- 소득(income)의 값이 일정할 때, 교육 수준의 한 단위 증가는 평균적으로 혜택 수준의 0.55 단위 증가와 연관이 있다.
- 교육 수준(education)의 값이 일정할 때, 소득의 한 단위 증가는 평균적으로 혜택 수준의 0.60 단위 증가와 연관이 있다.

그리고 상수인 절편(β_0) -6.1 은 다음과 같이 이해할 수 있음

- 만약 소득과 교육 수준 둘 모두가 0이라면, 혜택 수준의 예측값은 평균적으로 -6.1 이다.
- 그러나 과연 소득과 교육 수준이 모두 0인 경우가 현실에서 존재할까? → 굳이 β_0 를 열심히 해석하지 않는 이유

MLR coefficients

MLR: $\hat{Y} + \beta_0 + \beta_1 X_1 + \beta_2 X_2$ 가 있다고 할 때, 편미분을 통해 MLR의 각 예측변수의 부분 회귀계수를 구할 수 있음

- X_2 가 변하지 않는다고 할 때, X_1 의 Y 에 대한 부분적 효과: $\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$
- X_1 이 변하지 않는다고 할 때, X_2 의 Y 에 대한 부분적 효과: $\frac{\partial \hat{Y}}{\partial X_2} = \beta_2$

따라서 위의 Duncan의 자료를 이용한 회귀분석 결과를 해석해보면 다음과 같이 나타낼 수 있음

- 소득(income)의 값이 일정할 때, 교육 수준의 한 단위 증가는 평균적으로 혜택 수준의 0.55 단위 증가와 연관이 있다.
- 교육 수준(education)의 값이 일정할 때, 소득의 한 단위 증가는 평균적으로 혜택 수준의 0.60 단위 증가와 연관이 있다.

그리고 상수인 절편(β_0) -6.1은 다음과 같이 이해할 수 있음

- 만약 소득과 교육 수준 둘 모두가 0이라면, 혜택 수준의 예측값은 평균적으로 -6.1 이다.
- 그러나 과연 소득과 교육 수준이 모두 0인 경우가 현실에서 존재할까? $\rightarrow$ 굳이 β_0 를 열심히 해석하지 않는 이유

Several Explanatory Variables

k개의 변수를 가진 MLR

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon_i \\ &= \hat{Y}_i + \epsilon_i \end{aligned}$$

- SLR과 달리 여러 개의 변수를 가진 MLR의 경우, 모델을 시각화하여 보여주기는 어려움
- 그러나 잔차의 제곱합을 최소로 하는 계수값들을 구하는 것은 어려운 작업은 아님
  - $S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})]^2$
  - 이전 2개의 예측변수를 가졌던 MLR의 작업을 반복해주기만 하면 됨
  - 마찬가지로 모든 예측변수들 간 완벽한 선형 관계가 존재하거나 하나 이상의 예측변수가 변하지 않을 경우(invariant)를 제외하고는 각 계수값들은 단 하나의 해(solution)를 가질 것

Multiple Correlation

SLR과 마찬가지로 MLR에서의 잔차의 표준오차(the residual standard error)는 잔차의 '평균적인 크기'를 보여주는 통계치

- 단순하게 표본 크기(n)가 아니라 변수의 개수까지를 고려한 자유도(degree of freedom)에 따라서 $n-(k+1)=n-k-1$로 잔차의 분산을 나누어주면 잔차의 표준오차를 구할 수 있음

$$S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$

그리고 MLR의 잔차의 제곱합은 SLR과 동일한 방법으로 계산할 수 있음

- TSS $= \sum(Y_i - \bar{Y})^2$
- RegSS $= \sum(\hat{Y}_i - \bar{Y})^2$
- RSS $= \sum(Y_i - \hat{Y}_i)^2 = \sum E_i^2$

그리고 TSS $=$ RegSS $+$ RSS로 나타낼 수 있음

- OLS로 추정한 잔차는 종속변수의 예측값(fitted values)과 개별 예측변수와 독립적임

Multiple Correlation

따라서 MLR에서 다중상관(multiple correlation)의 제곱값(R^2)은

- 회귀분석을 이용했을 때, 회귀분석을 이용하지 않고 종속변수를 예측했을 때에 비해 얼마나 더 잘 예측했는지, 그 비율을 보여주는 값
- $R^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$
- 해석 자체는 SLR의 R^2와 크게 다르지 않음

그러나 R^2는 회귀분석 모델에 예측변수가 하나 추가될 때마다 오르기만 할 뿐, 내려가지는 않음

- 통계적으로 변수가 추가되었으면 종속변수를 어느 정도 설명하느냐의 차이가 있을지언정, 설명을 못하게 되지는 않음
- 따라서 R^2는 모델에 투입된 변수의 개수가 많으면 일단 높게 나옴
- 이를 보정하기 위해서 투입된 변수의 수, 자유도를 고려하여 조정된 R^2, $\tilde{R}^2$를 고안
  - $\tilde{R}^2 \equiv 1 - \frac{S_E^2}{S_Y^2} = 1 - \frac{\frac{\text{RSS}}{n-k-1}}{\frac{\text{TSS}}{n-1}}$
- 단, 표본의 규모가 매우 작을 경우, $\tilde{R}^2$는 R^2와 거의 다르지 않음

Standardized Regression Coefficients

사회과학 분야에서 변수들의 측정 단위가 다를 수 있음

- 이 경우 계수값의 직접적 비교가 어려울 수 있음
- 만약 표준화를 해준다면? 계수값의 크기들을 직관적으로 비교, 어떤 변수가 다른 변수에 비해 종속변수에 영향을 더/덜 미치는지 확인이 가능할 것
- 종속변수에 대한 예측변수들 간의 상대적 효과를 보여줌
- 하지만 어떻게 표준화를 할 것인가에 대한 이론적 근거는 존재하지 않음: 어디까지나 경험적인 접근
- 표준화에 관한 논의는 Gelman (2008) 참고

j 번째 예측변수에 대한 표준화된 부분회귀계수(Standardized partial regression coefficients)

- 다른 예측변수들의 값이 일정할 때, j 번째 예측변수가 1 표준편차 만큼 변화할 때, 평균적으로 나타나는 Y 값의 변화라고 이해할 수 있음
- 다만 실제적으로 이러한 표준화가 연구에 적합한지 고민하고 사용해야 함

Standardized Regression Coefficients

사회과학 분야에서 변수들의 측정 단위가 다를 수 있음

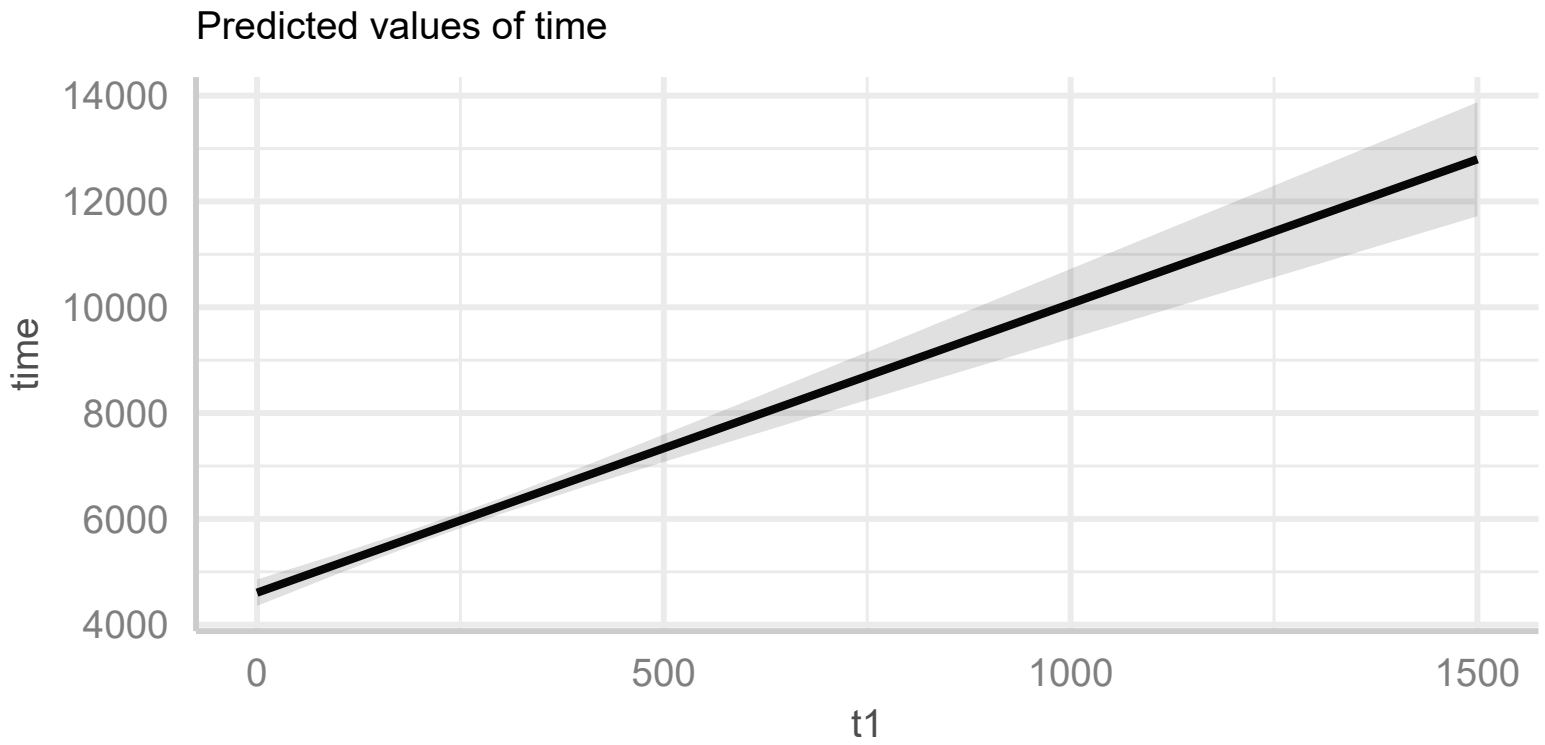
- 이 경우 계수값의 직접적 비교가 어려울 수 있음
- 만약 표준화를 해준다면? 계수값의 크기들을 직관적으로 비교, 어떤 변수가 다른 변수에 비해 종속변수에 영향을 더/덜 미치는지 확인이 가능할 것
- 종속변수에 대한 예측변수들 간의 상대적 효과를 보여줌
- 하지만 어떻게 표준화를 할 것인가에 대한 이론적 근거는 존재하지 않음: 어디까지나 경험적인 접근
- 표준화에 관한 논의는 Gelman (2008) 참고

j 번째 예측변수에 대한 표준화된 부분회귀계수(Standardized partial regression coefficients)

- 다른 예측변수들의 값이 일정할 때, j 번째 예측변수가 1 표준편차 만큼 변화할 때, 평균적으로 나타나는 Y 값의 변화라고 이해할 수 있음
- 다만 실제적으로 이러한 표준화가 연구에 적합한지 고민하고 사용해야 함

Visualization of MLR

`carData` 패키지의 `Transact` 데이터셋 사용 (Fox and Weisberg, 2019, 265)



가상으로 `t2`의 값이 일정할 때, `t1`의 변화가 예측값과 어떻게 대응하는지를 시각화로 나타낼 수 있음

Fox, John, and Sanford Weisberg. 2019. A R Companion to Applied Regression. Thrid ed. CA: SAGE Publications Ltd.

Fox, John Jr. 2016. Applied Regression Analysis and Generalized Linear Models. 3 ed. CA: Thousand Oaks, SAGE Publications.

Gelman, Andrew. 2008. “Scaling Regression Inputs by Dividing by Two Standard Deviations.” Statistics in Medicine 27: 2865–2873.

Table of Contents