KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

8. 단순회귀분석

정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

오늘의 목표

.pull-left[

10:05-10:45

중간시험에서 오답률이 높았던 문항들 위주로 검토

10:55-11:40

가설 검정의 논리에 대한 개요를 복습 및 학습

11:55-12:40

실습과제 해설 및 질의응답

보충과제 설명

Part I. 선형회귀의 기초

단순회귀분석

평균에서 관계로

우리가 데이터에서 얻고 싶은 가장 기본적인 정보는 무엇일까?

질문 1: Y 의 값은 무엇인가? (예: 한 반의 평균 키는?)

- 추가 정보 (X) 가 없을 때, Y 에 대한 최선의 예측은 **평균(Mean)**
- $Y_i = \mu + \epsilon_i$
- 모델 예측: $\hat{Y} = \bar{Y}$

단순회귀분석

평균에서 관계로

우리가 데이터에서 얻고 싶은 가장 기본적인 정보는 무엇일까?

질문 2: X 와 Y 는 관련이 있는가? (예: 키와 몸무게는 관련이 있는가?)

- 두 변수가 함께 움직이는지(방향, 강도)
- 이것이 바로 **공분산(Covariance)** 및 **상관계수(Correlation)**
- $r = \frac{S_{XY}}{S_X S_Y}$
- $-1 \leq r \leq 1$

단순회귀분석

상관관계의 한계

상관계수는 유용하지만, 우리가 원하는 모든 것을 알려주지는 않음.

상관관계는 단위가 없음(Unitless)

$r = 0.8$은 "강한 양의 관계"를 의미하지만, "X가 1단위 증가할 때 Y가 *얼마나* 증가하는지"는 알려주지 않음.

예측(Prediction)이 어려움

X 의 값을 알 때 Y 의 **구체적인 값**을 예측하는 명확한 모델을 제공하지 않음.

단순회귀분석

상관관계의 한계

인과관계(Causation)에 대한 정보를 제공하지 않음

상관관계는 대칭적($\mathrm{Cor}(X, Y) = \mathrm{Cor}(Y, X)$)

$X \to Y$ 인지 $Y \to X$ 인지, 혹은 제3의 변수 Z 때문인지 알 수 없음.

더 나은 질문: X가 1단위 변할 때, Y는 *평균적으로* 얼마나 변하는가?

$\to$ 이것이 바로 **회귀분석(Regression Analysis)**의 핵심 질문

단순회귀분석

기본적 구조

회귀분석은 X 를 기반으로 Y 를 예측하는 **선형 모델**을 적합

모집단 회귀 모델(Population Regression Model)

- 우리가 추정하고자 하는 실제 세계의(보이지 않는) 관계

$$Y_i = \alpha + \beta X_i + u_i$$

- Y_i: 종속변수(Dependent Variable)
- X_i: 설명변수(Independent Variable)
- α: 모집단의 절편(intercept)
- β: 모집단의 기울기(slope). X 가 1단위 증가할 때 Y 의 평균 변화량
- u_i: 모집단의 오차항(error term)

단순회귀분석

기본적 구조

표본 회귀 모델(Sample Regression Model)

우리가 가진 데이터(표본)를 사용해 모집단 모델을 *추정*한 것

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

- $\hat{\alpha}$: **추정된** 절편. 모집단 α의 추정치
- $\hat{\beta}$: **추정된** 기울기. 모집단 β의 추정치
- $\hat{u}_i$: **잔차(Residual)**. 모집단 오차 u_i의 추정치

단순회귀분석

기본적 구조

표본 회귀 모델(Sample Regression Model)

우리가 가진 데이터(표본)를 사용해 모집단 모델을 *추정*한 것

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$$

예측값(Fitted Value)

모델이 예측한 Y의 값

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

잔차(Residual)

실제 값과 예측값의 차이

$$\hat{u}_i = Y_i - \hat{Y}_i$$

단순회귀분석

기본적 구조

최적의 선이란 무엇인가?

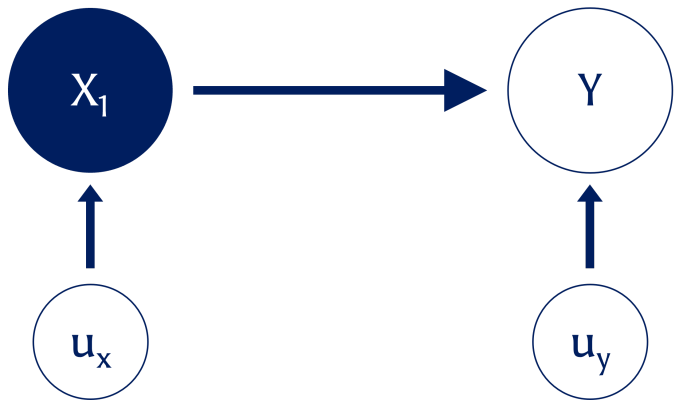
- 우리의 목표: 실제 데이터(Y_i)와 모델 예측값($\hat{Y}_i$)의 차이, 즉 **잔차($\hat{u}_i$)를 가능한 한 작게** 만드는 $\hat{\alpha}$와 $\hat{\beta}$를 찾는 것

문제점: 그냥 잔차를 모두 더하면 어떻게 될까?

- $\sum \hat{u}_i = \sum (Y_i - \hat{Y}_i)$
- OLS 추정의 수학적 속성에 의해, 잔차의 합은 항상 **0**이 됨: $\sum \hat{u}_i = 0$
- 양의 잔차(+5)와 음의 잔차(-5)가 서로를 상쇄시켜, 모델이 얼마나 "나쁜지" 비교할 수 없음.

단순회귀분석

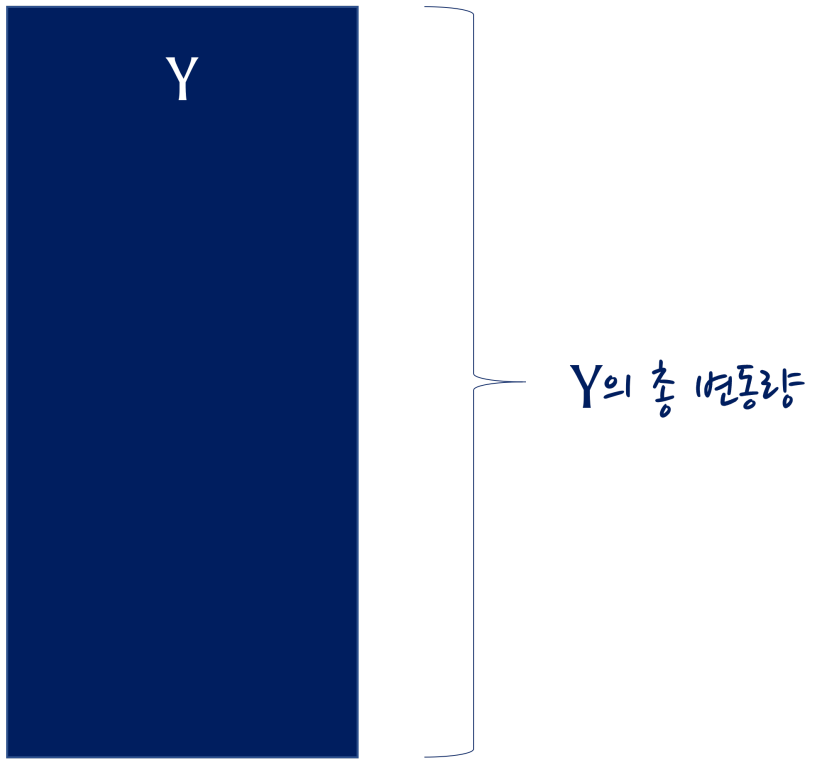
개념도



- Y: 종속변수
- X_1: 예측변수
- u_x: 예측변수에 영향을 미칠 수 있는 외부의 요인
- u_y: 종속변수에 영향을 미칠 수 있는 외부의 요인
- 관계 양상은 X_1으로부터 Y, 즉, X_1이 Y의 원인

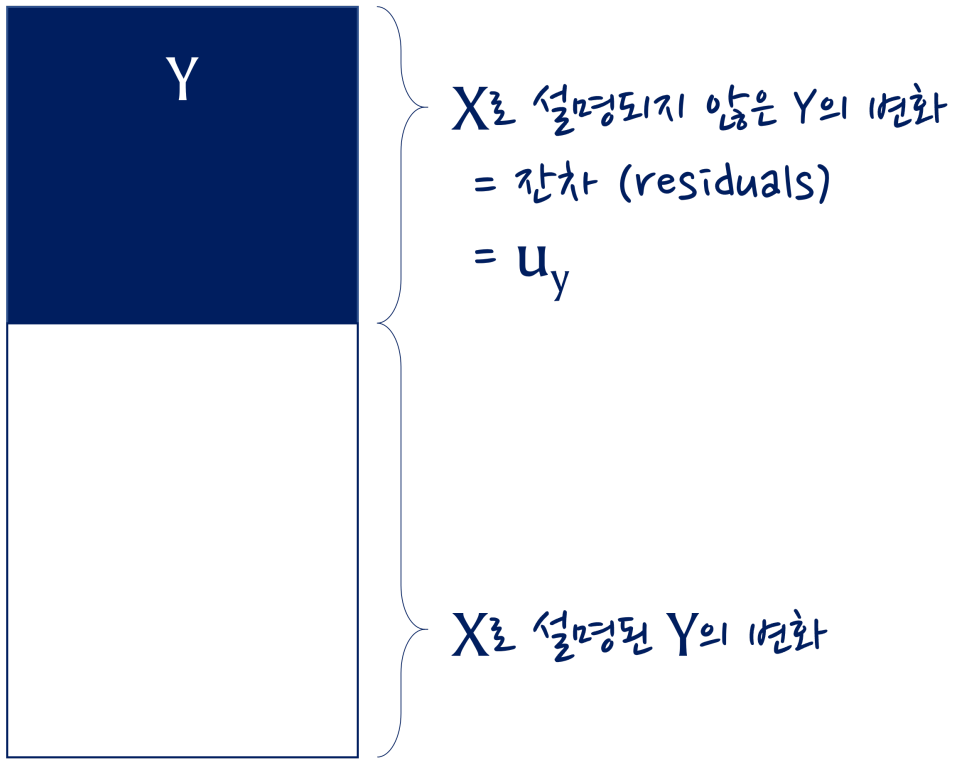
단순회귀분석

개념도



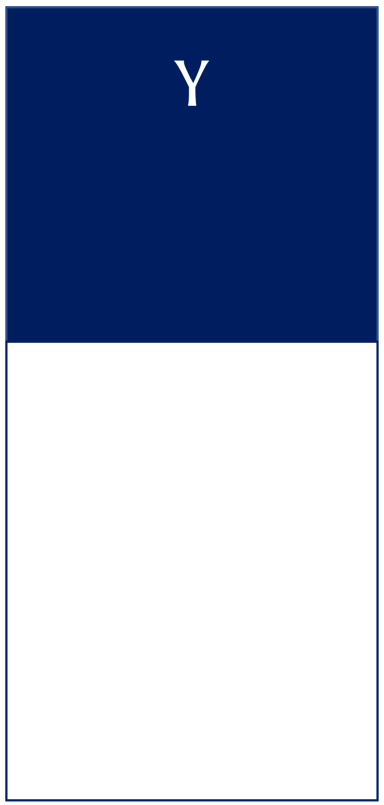
단순회귀분석

개념도



단순회귀분석

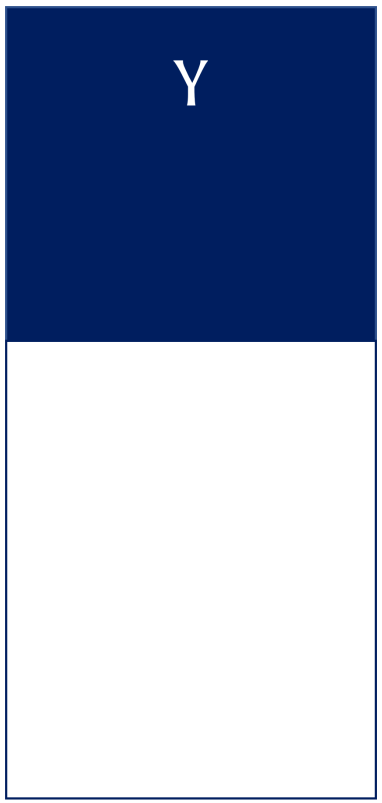
개념도



$$\frac{\text{Y의 총 변화량}}{\text{X로 설명된 Y의 변화}} = R^2$$

단순회귀분석

개념도



Y의 총 변화량

X로 설명된 Y의 변화

=

모델이 Y의 변화량의 몇 %를 설명하는지?

단순회귀분석

최소제곱기준(Least-Squares Criterion)

부호 문제를 해결하고 잔차의 "총 크기"를 측정하기 위해, 잔차를 **제곱**하여 총합을 구함.

잔차제곱합 (Residual Sum of Squares; RSS)

$$S(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - (\hat{\alpha} + \hat{\beta} X_i))^2$$

최소제곱법 (Ordinary Least Squares; OLS)

RSS를 **최소화**하는 $\hat{\alpha}$와 $\hat{\beta}$를 찾는 방법

"Ordinary" (보통) : 가장 기본적이고 널리 쓰이기 때문

단순회귀분석

OLS 추정치 찾기: (1) 미적분

RSS 함수 $S(\hat{\alpha}, \hat{\beta})$를 최소화하는 값을 찾기 위해, 각 모수에 대해 **편미분(partial derivatives)**을 하고 그 값을 0으로 둠.

1. $\hat{\alpha}$에 대해 편미분
   - $\frac{\partial S(\hat{\alpha},\hat{\beta})}{\partial \hat{\alpha}} = \sum(2)(Y_i - \hat{\alpha} - \hat{\beta}X_i)(-1) = 0$
   - $\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i$
   - $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X} \implies \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$

단순회귀분석

OLS 추정치 찾기: (1) 미적분

RSS 함수 $S(\hat{\alpha}, \hat{\beta})$를 최소화하는 값을 찾기 위해, 각 모수에 대해 **편미분(partial derivatives)**을 하고 그 값을 0으로 둠.

1. $\hat{\beta}$에 대해 편미분
   - $\frac{\partial S(\hat{\alpha},\hat{\beta})}{\partial \hat{\beta}} = \sum(2)(Y_i - \hat{\alpha} - \hat{\beta}X_i)(-X_i) = 0$
   - $\sum X_i Y_i = \hat{\alpha}\sum X_i + \hat{\beta}\sum X_i^2$

두 식을 연립하여 풀면 $\hat{\beta}$의 공식을 얻게 됨.

단순회귀부니석

OLS 추정치 공식

미적분을 통해 유도된 $\hat{\alpha}$와 $\hat{\beta}$의 공식은 다음과 같음.

기울기 (Slope)

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- 분자: X 와 Y 의 **공분산(Covariance)**, X 와 Y 가 함께 움직이는 정도, 즉 관계의 방향성과 강도
- 분모: X 의 **분산(Variance)**, X 값들이 평균에서 얼마나 퍼져 있는지를 보여줌.

X 가 한 단위 변할 때 Y 가 평균적으로 얼마나 변하는지를 나타냄. X 의 변동이 Y 의 변동을 얼마나 "설명"하는지 보여주는 비율

단순회귀분석

OLS 추정치 공식

절편 (Intercept)

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

- 이 식은 OLS 회귀선이 항상 **평균 지점** $(\bar{X}, \bar{Y})$ 을 통과한다는 것을 의미

단순회귀분석

OLS 추정치 찾기: (2) 행렬 대수

단순회귀분석은 간단하지만, X가 10개가 되면 미적분은 매우 복잡해짐.

행렬(Matrix)을 사용하면 다중회귀분석까지 동일한 공식으로 표현할 수 있음.

모든 관측치를 행렬로 표현: $Y = X\beta + u$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

단순회귀분석

OLS 추정치 찾기: (2) 행렬 대수

단순회귀분석은 간단하지만, X가 10개가 되면 미적분은 매우 복잡해짐.

행렬(Matrix)을 사용하면 다중회귀분석까지 동일한 공식으로 표현할 수 있음.

모든 관측치를 행렬로 표현: $Y = X\beta + u$

- Y: 종속변수 벡터 ($n \times 1$)
- X: 독립변수 행렬 ($n \times (k+1)$, 단순회귀는 $n \times 2$)
- β: 계수 벡터 ($(k+1) \times 1$, 단순회귀는 2×1)
- u: 오차항 벡터 ($n \times 1$)

단순회귀분석

OLS 행렬 공식

목표: 잔차제곱합 $\sum \hat{u}_i^2 = \hat{u}'\hat{u}$ 를 최소화

- $\hat{u} = Y - X\hat{\beta}$
- $RSS = (Y - X\hat{\beta})'(Y - X\hat{\beta})$

행렬 미적분을 통해 이 RSS를 최소화하는 $\hat{\beta}$를 풀 수 있음.

단순회귀분석

OLS 행렬 공식

OLS 추정치의 행렬 공식

$$\hat{\beta} = (X'X)^{-1}X'Y$$

- $X'X$: (2×2) 행렬 $\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$
- $X'Y$: (2×1) 벡터 $\begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}$

이 공식을 풀면 미적분으로 구한 $\hat{\alpha}$와 $\hat{\beta}$ 값과 정확히 일치

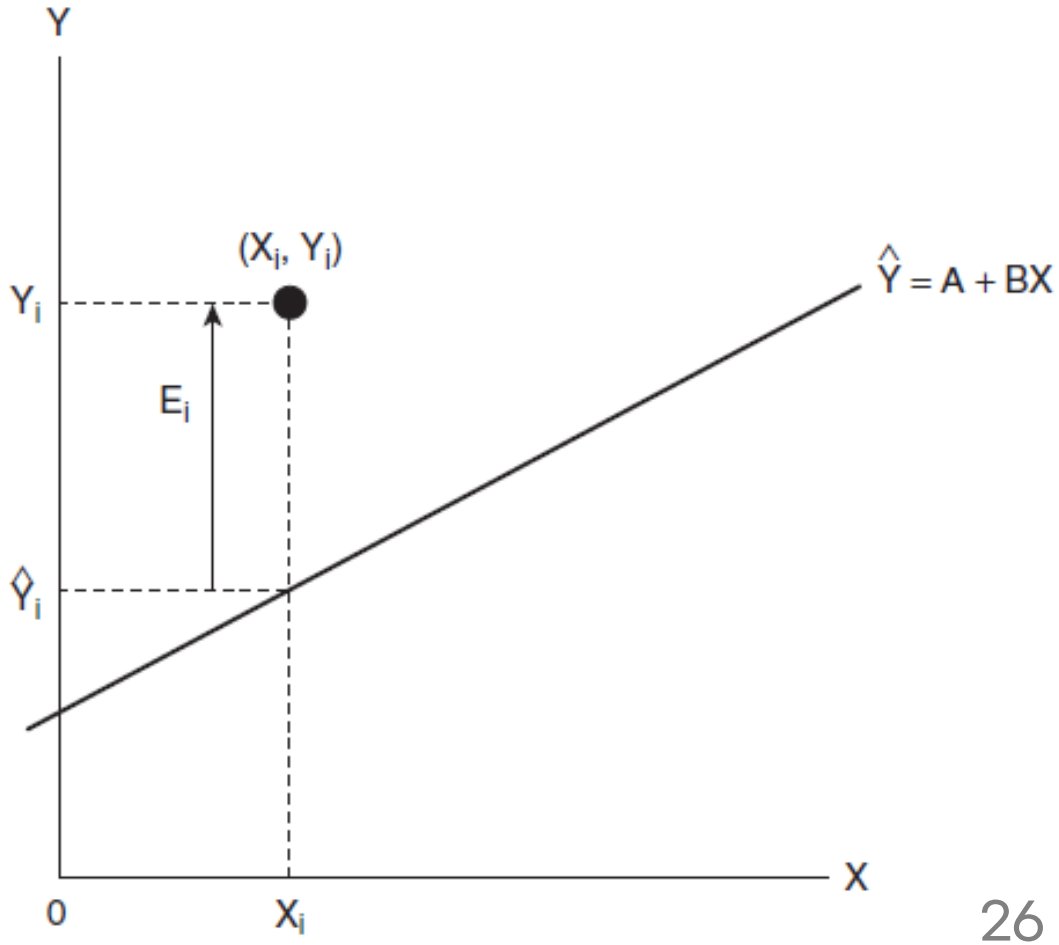
이 공식은 X 변수가 100개인 **다중회귀분석**에서도 동일하게 작동

단순회귀분석

예제: 계수 해석하기

$$E_i = Y_i - \hat{Y}_i = Y_i - (A + BX_i)$$

- i 번째 관측치에 대한 잔차(E_i)를 보여주는 Y 에 대한 X 의 회귀곡선
- 잔차(E_i)란 실제 개별 관측치(Y_i)로부터 모델로 예측한 값($\hat{Y}_i$)의 차이



단순회귀분석

예제: 계수 해석하기

$\widehat{\text{측정 몸무게}} = 1.78 + 0.977 \times \text{보고 몸무게}$

- $\hat{\beta} = 0.977$: "보고된 몸무게가 1kg 증가할 때, 측정된 몸무게는 **평균적으로** 약 0.977 kg 증가와 관련이 있다."
  - "관련이 있다" (associated with) $\neq$ "원인이다" (causes)
- $\hat{\alpha} = 1.78$: $X = 0$ (보고 몸무게=0)일 때의 Y (측정 몸무게) 예측값
  - 현실적으로 몸무게가 0kg일 수 없으므로, 이 값은 **수학적 의미**는 있으나 **실질적 해석**에는 의미가 없을 수 있음.
- 만약 보고된 몸무게가 완벽히 정확했다면 $\hat{\alpha} = 0, \hat{\beta} = 1$이었을 것임.

Part II. 모형 적합도

단순회귀분석

모형 적합도(Goodness-of-Fit)

모델을 만들었다면, 이 모델이 데이터를 얼마나 잘 설명하는지 평가해야 함.

1. R^2 **(결정계수):** 상대적 적합도
2. **Root MSE (RMSE):** 절대적 적합도

단순회귀분석

모형 적합도(Goodness-of-Fit)

TSS (Total Sum of Squares): 총 변동량

- X 없이 Y를 평균($\bar{Y}$)으로 예측할 때의 총 오차
- $TSS = \sum(Y_i - \bar{Y})^2$

RSS (Residual Sum of Squares): 잔차 변동량

- X를 사용한 회귀모델로 예측할 때의 남은 오차
- $RSS = \sum(Y_i - \hat{Y}_i)^2 = \sum \hat{u}_i^2$

단순회귀분석

모형 적합도: R^2 (R-squared)

R^2 (결정계수): 단순회귀에서는 $R^2 = r^2$ (상관계수의 제곱)

- Y의 총 변동량(TSS) 중, 우리 모델(X)이 **설명해낸 변동량의 비율**
- $0 \leq R^2 \leq 1$

$$R^2 \equiv \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- $R^2 = 1$: 완벽한 설명. 모든 점이 회귀선 위에 있음 (RSS=0).
- $R^2 = 0.32$: "우리 모델(X)이 Y의 전체 변동 중 32%를 설명한다."
- $R^2 = 0$: 완전한 무설명. 모델이 $\bar{Y}$로 예측하는 것보다 나을 게 없음 (RSS=TSS).

단순회귀분석

모형 적합도: RMSE

RMSE (Root Mean Squared Error)

- "회귀의 표준오차" (Standard Error of the Regression) 또는 "잔차의 표준편차" (S_E)라고도 불림.
- 모델의 예측이 **평균적으로 얼마나 빗나가는지**를 *종속변수(Y)와 동일한 단위*로 보여줌.

$$RMSE = S_E = \sqrt{\frac{\sum \hat{u}_i^2}{n-k-1}} = \sqrt{\frac{RSS}{n-2}}$$

- n: 관측치 수
- k: 독립변수 수 (단순회귀는 $k=1$)
- $n-k-1$: 자유도 (degrees of freedom)

단순회귀분석

모형 적합도: RMSE

RMSE (Root Mean Squared Error)

해석:

- "RMSE = 4.83"
- "우리 모델은 Y를 예측할 때 평균적으로 ± 4.83 포인트의 오차를 가진다."
- R^2와 달리 RMSE는 낮을수록 좋다고 할 수 있음.

단순회귀분석

OLS의 가정(Gauss-Markov Assumptions)

OLS가 "좋은" 추정치(예: 편향되지 않음)가 되기 위해서는 특정 가정들이 필요

Gauss-Markov 5대 가정 이 가정 충족 시 OLS는 BLUE가 됨.

선형성(Linearity): 모집단 모형이 모수(α, β)에 대해 선형

- $Y = \alpha + \beta X + u$

무작위 표본(Random Sample): 데이터가 모집단에서 무작위로 추출

단순회귀분석

OLS의 가정(Gauss-Markov Assumptions)

OLS가 "좋은" 추정치(예: 편향되지 않음)가 되기 위해서는 특정 가정들이 필요

Gauss-Markov 5대 가정 이 가정 충족 시 OLS는 BLUE가 됨.

외생성(Exogeneity / Zero Conditional Mean)

- $E(u_i|X_i) = 0$
- **가장 중요한 가정.** X의 값에 상관없이, 오차항의 평균은 0이다.
  - X와 u가 상관관계가 없어야 함 ($\mathrm{Cov}(X, u) = 0$).
- *위반 시:* **내생성(Endogeneity)** 문제 발생. $\hat{\beta}$가 **편향(biased)**됨.
  - 누락변수 편향 - Omitted Variable Bias

단순회귀분석

OLS의 가정(Gauss-Markov Assumptions)

OLS가 "좋은" 추정치(예: 편향되지 않음)가 되기 위해서는 특정 가정들이 필요

Gauss-Markov 5대 가정 이 가정 충족 시 OLS는 BLUE가 됨.

등분산성(Homoskedasticity)

- X의 값에 관계없이 오차항(u_i)의 분산이 일정
- $Var(u_i|X_i) = \sigma^2$ (일정한 상수)
  - *위반 시:* **이분산성(Heteroskedasticity)**
  - $\hat{\beta}$는 여전히 편향되지 않지만, **비효율적(inefficient)**이 되며, **표준오차(SE) 추정이 부정확**해져 $t-$검정이 망가짐.

단순회귀분석

OLS의 가정(Gauss-Markov Assumptions)

OLS가 "좋은" 추정치(예: 편향되지 않음)가 되기 위해서는 특정 가정들이 필요

Gauss-Markov 5대 가정 이 가정 충족 시 OLS는 BLUE가 됨.

완전 공선성 부재(No Perfect Collinearity)

- X 변수가 상수가 아니어야 함($Var(X) > 0$).
    - 다중회귀에서는: 어떤 X도 다른 X들의 선형조합이 아니어야 함.
    - *위반 시:* $\hat{\beta}$를 계산(행렬의 역행렬)할 수 없음.

단순회귀분석

OLS의 가정(Gauss-Markov Assumptions)

BLUE 란 무엇인가?

Gauss-Markov 정리: 위의 **5가지 가정이 모두 충족**될 때, OLS 추정치($\hat{\beta}$)는 **BLUE**

- **B**est (최선): 가장 효율적(efficient). 즉, **가장 작은 분산**을 가짐.
- **L**inear (선형): Y의 선형 결합으로 계산됨 ($\hat{\beta} = \sum w_i Y_i$).
- **U**nbiased (비편향): 평균적으로 $\hat{\beta}$는 실제 β와 같다. $E(\hat{\beta}) = \beta$.
- **E**stimator (추정량): 모집단의 모수를 추정하는 공식.

즉, OLS는 (가정이 맞다면) 우리가 사용할 수 있는 수많은 "선형 비편향 추정량" 중에서 가장 정확하고 신뢰할 수 있는(분산이 작은) 추정량

단순회귀분석

추론(Inference)을 위한 추가 가정

정규성 가정 (Normality Assumption)

- 오차항 u_i가 정규분포를 따른다: $u_i \sim N(0, \sigma^2)$

이 가정이 왜 필요한가?

- 이 가정은 **BLUE**가 되기 위해 *필요하지 않음.* 이 가정은 **작은 표본(small n)**에서 **추론(Inference)**을 할 때 필요
  - 즉, $\hat{\beta}$의 분포가 t-분포를 따른다고 가정할 수 있게 해줌.
  - p-값, t-통계량, 신뢰구간(CI) 계산이 가능해짐.
- 표본이 매우 크다면($n \approx 100+$), **중심극한정리(CLT)**에 의해 u_i의 분포와 상관없이 $\hat{\beta}$의 표집분포가 정규분포에 근사하므로, 이 가정은 덜 중요해짐.

Part III. 단순선형회귀 R 코드 실습 및 실습과제 9 설명

단순회귀분석

R 예제

```
# 필요한 패키지
library(tidyverse); library(moderndive); library(skimr); library(gapminder)

evals_ch5 <- evals |> dplyr::select(ID, score, bty_avg, age)

head(evals_ch5)
```

```
## # A tibble: 6 × 4
##      ID score bty_avg   age
##   <int> <dbl>   <dbl> <int>
## 1     1   4.7       5    36
## 2     2   4.1       5    36
## 3     3   3.9       5    36
## 4     4   4.8       5    36
## 5     5   4.6       3    59
## 6     6   4.3       3    59
```

단순회귀분석

R 예제

```
evals_ch5 |>
  summarize(mean_bty_avg = mean(bty_avg), mean_score = mean(score),
            median_bty_avg = median(bty_avg), median_score = median(score))
```

```
## # A tibble: 1 × 4
##   mean_bty_avg mean_score median_bty_avg median_score
##          <dbl>      <dbl>          <dbl>        <dbl>
## 1         4.42       4.17           4.33          4.3
```

```
evals_ch5 |>
  get_correlation(formula = score ~ bty_avg)
```

```
## # A tibble: 1 × 1
##     cor
##   <dbl>
## 1 0.187
```

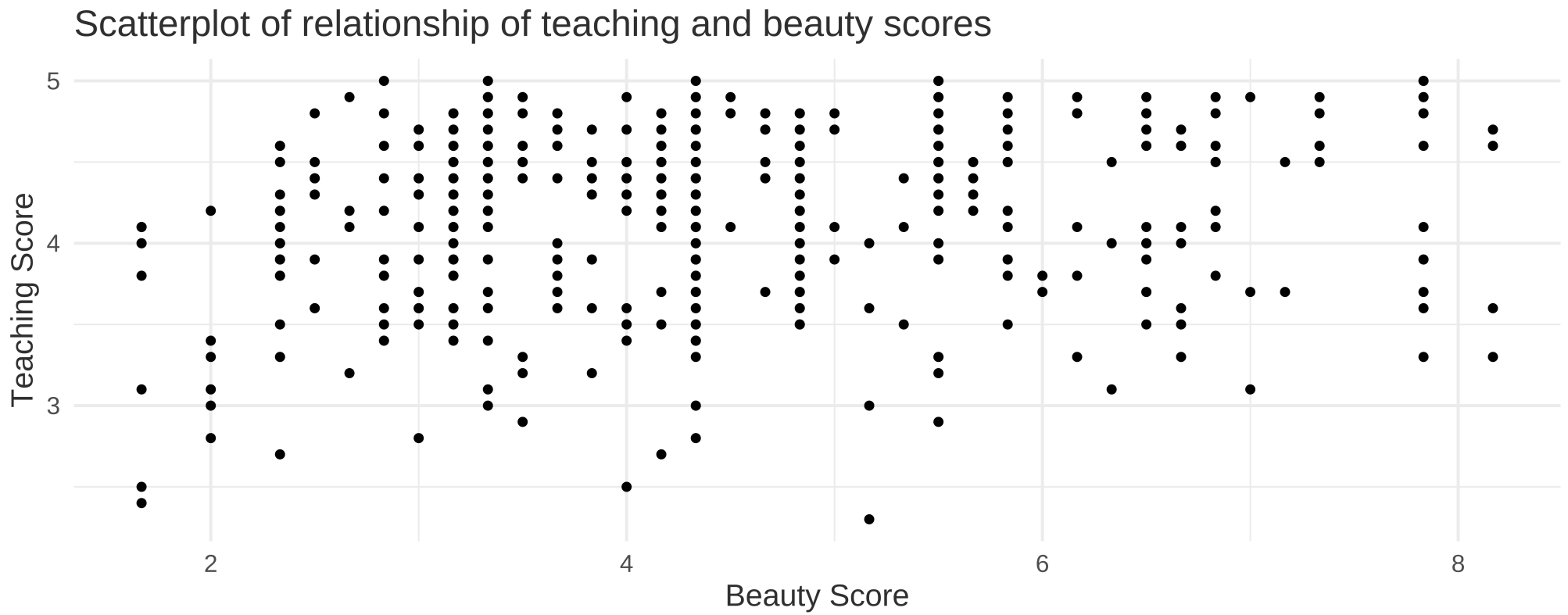
단순회귀분석

R 예제

```
evals_ch5 |>
  ggplot(aes(x = bty_avg, y = score)) +
  geom_point() +
  labs(x = "Beauty Score",
       y = "Teaching Score",
       title = "Scatterplot of relationship of teaching and beauty scores")
```

단순회귀분석

R 예제


Scatterplot of relationship of teaching and beauty scores
Teaching Score
5
4
3
2
4
6
8
Beauty Score

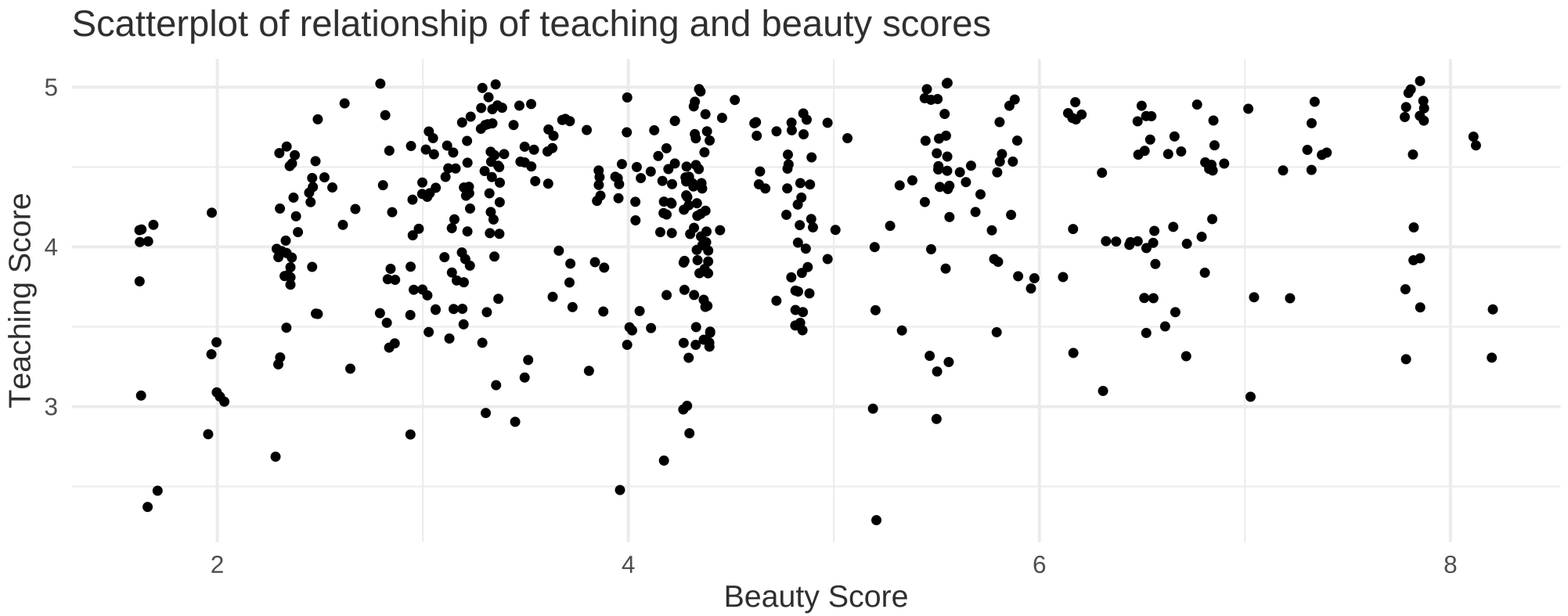
단순회귀분석

R 예제

```
evals_ch5 |> ggplot(aes(x = bty_avg, y = score)) +
  geom_jitter() +
  labs(x = "Beauty Score", y = "Teaching Score",
       title = "Scatterplot of relationship of teaching and beauty scores")
```

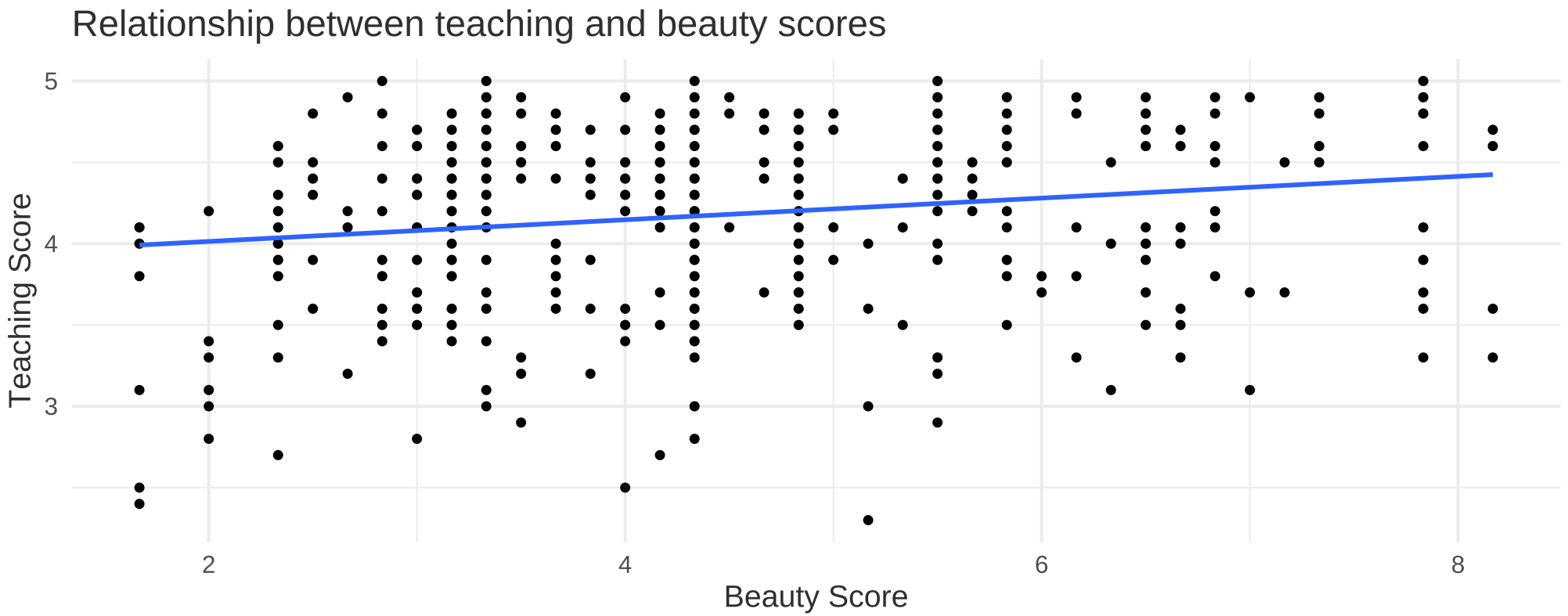
단순회귀분석

R 예제


Scatterplot of relationship of teaching and beauty scores
Teaching Score
5
4
3
2
4
6
8
Beauty Score

단순회귀분석

R 예제


Relationship between teaching and beauty scores
Teaching Score
5
4
3
2
4
6
8
Beauty Score

단순회귀분석

R 예제

```
# Fit regression model:
score_model <- lm(score ~ bty_avg, data = evals_ch5)
# Get regression table:
get_regression_table(score_model)
```

```
## # A tibble: 2 × 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    3.88      0.076     51.0        0    3.73     4.03
## 2 bty_avg      0.067     0.016      4.09       0    0.035    0.099
```

$$\widehat{\text{Score}} = 0.067 \times (\text{Average Beauty Score}) + 3.880.$$

$$\hat{\beta} : 0.067, \;\; \alpha : 3.880, \;\; \epsilon : \text{Score} - \widehat{\text{Score}}$$

KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

감사합니다!

궁금한 것이 있으면 언제든 연락하세요.

강사 연락처

연락처	박상훈
	sh.park.poli@gmail.com
	sanghoon-park.com/
	영상바이오관 405