KNU 강원대학교
KANGWON NATIONAL UNIVERSITY

12. 비선형 관계와 로지스틱 회귀분석

정치와 데이터분석

박상훈 (sh.park.poli@gmail.com)
강원대학교

오늘의 목표

종속변수가 0, 1일 때?

선형확률모형(LPM)의 적용과 한계

- 이항 변수(0/1)에 OLS를 적용했을 때의 문제점 이해

최대우도추정(MLE)의 개념적 이해

로지스틱 회귀분석의 기초: 승산(Odds)과 로짓(Logit) 변환의 이해

로지스틱 모형의 해석과 Log-odds 계수의 해석

승산비(Odds Ratios)의 해석

모형의 평가

분류 정확도와 ROC 곡선

Part I. 종속변수가 0, 1일 때, 우리의 선택?

비선형 관계와 로지스틱 회귀분석

선형확률모형(Linear Probability Model; LPM)

종속변수가 이항변수(Binary)일 때 정치학이나 사회과학 데이터에서는 종속변수가 0 또는 1인 경우가 매우 많음.

예시:

- 투표 여부 (투표함 = 1, 기권 = 0)
- 선거 당선 여부 (당선 = 1, 낙선 = 0)
- 전쟁 발발 여부 (발발 = 1, 평화 = 0)
- 법안 통과 여부 (가결 = 1, 부결 = 0)

이러한 데이터를 분석할 때, 우리가 기존에 배운 **OLS(최소자승법)**를 그대로 사용하면 어떤 문제가 발생할까?

비선형 관계와 로지스틱 회귀분석

선형확률모형(Linear Probability Model; LPM)

이항 종속변수 Y 에 대해 OLS를 적용하는 것을 선형확률모형(LPM)이라고 함.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

여기서 $\mathbb{E}(Y_i|X_i)$ 는 X_i 가 주어졌을 때, Y_i 가 1일 확률, 즉 $\Pr(Y_i = 1)$ 로 해석됨.

예제 데이터 생성 | LPM 적합 및 시각화

```
set.seed(2024)
# 가상의 유권자 데이터 생성 (나이와 투표여부)
data <- tibble(
  age = runif(100, 18, 90),
  # 나이가 많을수록 투표 확률 증가하도록 설정 (로지스틱 함수 이용)
  prob = 1 / (1 + exp(-(-5 + 0.1 * age))),
  vote = rbinom(100, 1, prob)
)
```

비선형 관계와 로지스틱 회귀분석

선형확률모형(Linear Probability Model; LPM)

문제점 1: 확률의 범위 위반

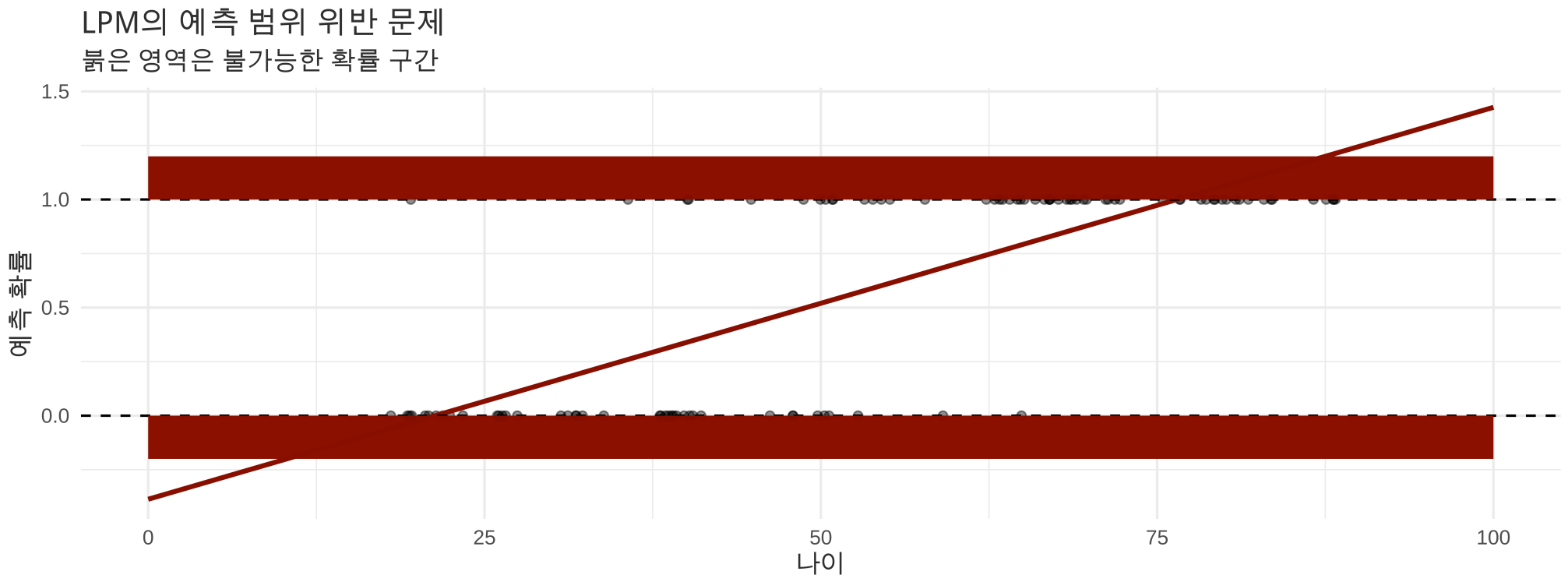
확률은 0과 1사이여야 함 $(0 \leq \Pr \leq 1)$.

그러나 LPM은 선형 $(Y = a + bX)$ 이기 때문에, X 값이 매우 작거나 크면 **예측값이 0보다 작거나 1보다 커질 수 있음.**

비선형 관계와 로지스틱 회귀분석

선형확률모형(Linear Probability Model; LPM)

문제점 1: 확률의 범위 위반



비선형 관계와 로지스틱 회귀분석

선형확률모형(Linear Probability Model; LPM)

문제점 2: 오차항의 비정규성과 이분산성

비정규성(Non-normality)

- Y 는 0 또는 1이므로, 오차항 $\epsilon = Y - \hat{Y}$ 는 두 가지 값만 가질 수 있음.
- 따라서 오차항이 정규분포를 따른다는 OLS 가정이 위배됨.

이분산성(Heteroskedasticity)

- 베르누이 분포의 분산은 $\Pr(1 - \Pr)$ 임.
- $\Pr$ 가 X 에 따라 변하므로, 분산도 X에 따라 변함. 즉, 분산이 일정하지 않음.
    - Gauss-Markov 가정 위배 $\rightarrow$ 표준오차 추정치 편향

비선형 관계와 로지스틱 회귀분석

선형확률모형(Linear Probability Model; LPM)

문제점 3: 비현실적인 선형 가정

현실 세계에서 확률은 선형적으로 증가하지 않는 경우가 많음.

- 예: 나이가 20세에서 21세로 늘 때 투표 확률 증가폭과, 80세에서 81세로 늘 때의 증가폭은 다를 수 있음.
- 보통 **S자 형태(Sigmoid)**의 곡선을 그리는 경우가 많음(어느 임계점까지 서서히 증가하다 급격히 증가하고, 다시 완만해짐).

그래서 우리는 **로지스틱 회귀분석(Logistic Regression)**을 사용한다!

Part II. 로지스틱 회귀분석의 구조

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

일반화 선형 모형(Generalized Linear Models, GLM)

로지스틱 회귀모형은 GLM의 일종

GLM은 세 가지 요소로 구성

1. 랜덤 성분(Random Component): 종속변수 Y 의 확률분포(예: 이항분포, 정규분포, 포아송분포 등)
2. 선형 예측자(Linear Predictor): 설명변수들의 선형 결합 $(\eta_i = \beta_0 + \beta_1 X_{1i} + \dots)$
3. 연결 함수(Link Function): 선형 예측자 (η)와 종속변수의 평균 (μ)을 연결하는 함수 $(g(\mu) = \eta)$

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

일반화 선형 모형(Generalized Linear Models, GLM)

로지스틱 회귀모형은 GLM의 일종

GLM은 세 가지 요소로 구성

로지스틱 회귀의 경우, 랜덤 성분은 이항 분포(Binomial Distribution)이며, 연결 함수는 로짓 함수(Logit Function)인 셈

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

일반화 선형 모형(Generalized Linear Models, GLM)

연결 함수가 필요한 이유

우리는 선형 예측자 $\beta_0 + \beta_1 X$ 를 사용하여 확률 π (0 ~ 1 사이의 값)를 예측하고 싶음.

하지만 선형식은 $-\infty$ 에서 $+\infty$ 까지 범위 $\rightarrow$ **범위를 [0, 1]로 가두기 위해 변환이 필요함.**

가장 대표적인 변환 함수가 **로지스틱 함수(Logistic function)**와 **정규 누적분포함수 (Probit)**

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

승산(Odds)

확률(Probability) $\Pr$ 가 사건이 일어날 가능성이라면, **승산(Odds)**은 (일어날 확률 / 일어나지 않을 확률)의 비율

$$Odds = \frac{\Pr}{1-\Pr}$$

- $\Pr = 0.5$ (반반) $\rightarrow Odds = 1$
- $\Pr = 0.8 \rightarrow Odds = 4\ (0.8/0.2)$
- $\Pr = 0.2 \rightarrow Odds = 0.25\ (0.2/0.8)$

Odds의 범위는 0 에서 $+\infty$ 까지임.

여전히 음수를 가질 수 없음.

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

로지스틱 회귀 모형식

따라서 로지스틱 회귀 모형은 다음과 같음:

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X$$

이를 확률 $\Pr$ 에 대해 정리하면 (역함수)

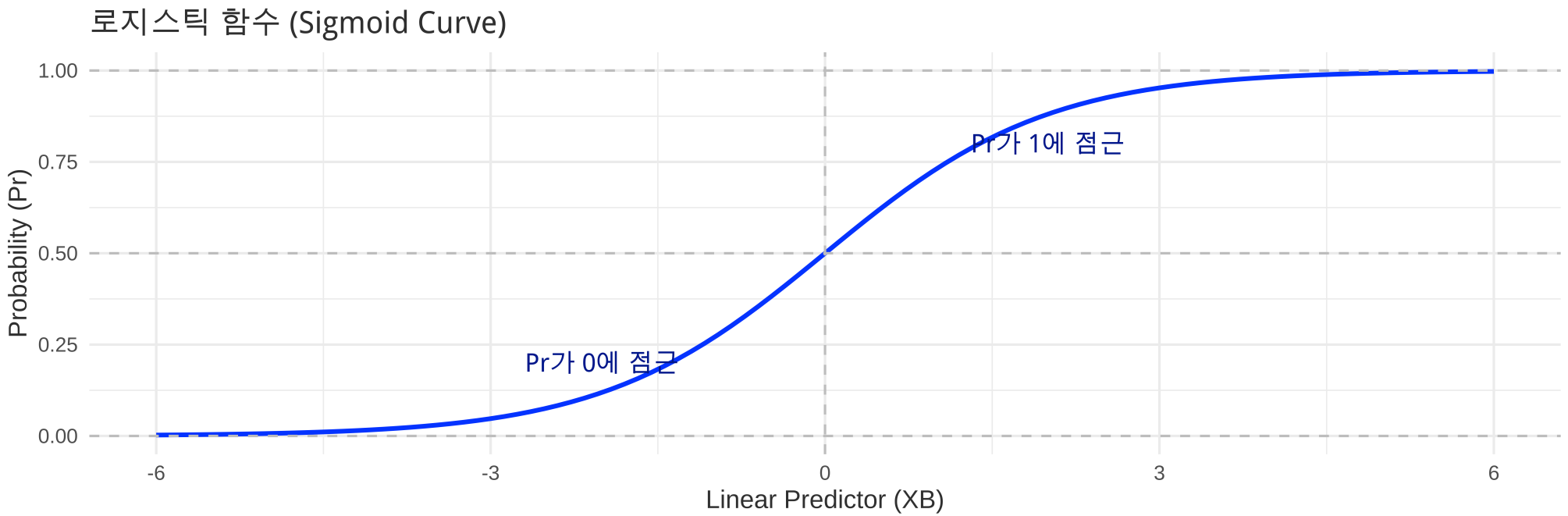
$$\Pr(Y=1) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

이것이 바로 **S자 곡선(Sigmoid Curve)**을 만드는 수식

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

로지스틱 회귀 모형식



비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

추정 방법: 최대가능도추정(MLE)

OLS는 오차의 제곱합(SSE)을 최소화하지만, 로지스틱 회귀는 **최대가능도추정(Maximum Likelihood Estimation, MLE)**을 사용

가능도(Likelihood)란? 관측된 데이터가 주어졌을 때, 특정 모수 (β) 값이 이 데이터를 만들어냈을 '가능성'

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

추정 방법: 최대가능도추정(MLE)

MLE의 목표: 우리가 관측한 데이터(예: 투표함, 투표안함)가 나올 확률을 최대화하는 β 값을 찾자.

$$L(\beta) = \prod_{i=1}^{n} P(Y_i = 1)^{y_i} (1 - P(Y_i = 1))^{1-y_i}$$

계산의 편의를 위해 양변에 로그를 취한 **로그 우도(Log-Likelihood)**를 최대화

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

계수, 승산비, 한계효과

R에서는 `glm()` (generalized linear model) 함수를 사용함. 로지스틱의 경우에는 `family = binomial(link = "logit")` 옵션이 필수

```
# 위에서 생성한 가상 데이터(data) 사용
# 종속변수: vote (0/1), 독립변수: age

logit_model <-
  glm(vote ~ age, data = data, family = b

texreg::screenreg(logit_model, single.row
```

```
##
## ================================
##                   Model 1
## --------------------------------
## (Intercept)      -7.01 (1.40) ***
## age               0.15 (0.03) ***
## --------------------------------
## AIC              63.30
## BIC              68.51
## Log Likelihood  -29.65
## Deviance         59.30
## Num. obs.       100
## ================================
```

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

계수, 승산비, 한계효과

Log-Odds 계수

위 결과에서 age의 계수(Estimate)를 보자.

- 계수 (β) : 독립변수 X 가 1단위 증가할 때, **로그 승산(Log-Odds)**의 변화량
- 수치 그 자체를 직관적으로 해석하기는 매우 어려움.
    - 부호(+): 나이가 들수록 투표할 확률(log-odds)이 증가한다.
    - 부호(-): 나이가 들수록 투표할 확률이 감소한다.
    - $p-$value: 통계적으로 유의미한가?

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

승산비(Odds Ratios)

계수 β 를 지수화 (e^{β}) 하면 **승산비(Odds Ratio, OR)**가 됨.

$$\frac{Odds(X+1)}{Odds(X)} = e^{\beta}$$

- $OR > 1$: 성공 확률(Odds) 증가
- $OR < 1$: 성공 확률(Odds) 감소
- $OR = 1$: 변화 없음

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

승산비(Odds Ratios)

```
library(broom)
# 계수 추출 및 지수화
coef_table <- tidy(logit_model) |>
  mutate(OR = exp(estimate)) |> # 승산비 계산
  select(term, estimate, OR, p.value)

coef_table
```

```
## # A tibble: 2 × 4
##   term        estimate       OR     p.value
##   <chr>          <dbl>    <dbl>       <dbl>
## 1 (Intercept)   -7.01  0.000902 0.000000599
## 2 age            0.146 1.16     0.000000197
```

나이가 1살 증가할 때, 투표할 **승산(Odds)**은 약 1.16배 증가한다.

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

예측 확률(Predicted Probabilities)

가장 직관적인 방법은 특정 X 값일 때의 예측 확률을 구하는 것

`predict()` 함수 사용:

- `type = "link"` : Logit 값 반환 (기본값)
- `type = "response"` : 확률 값 (0~1) 반환

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

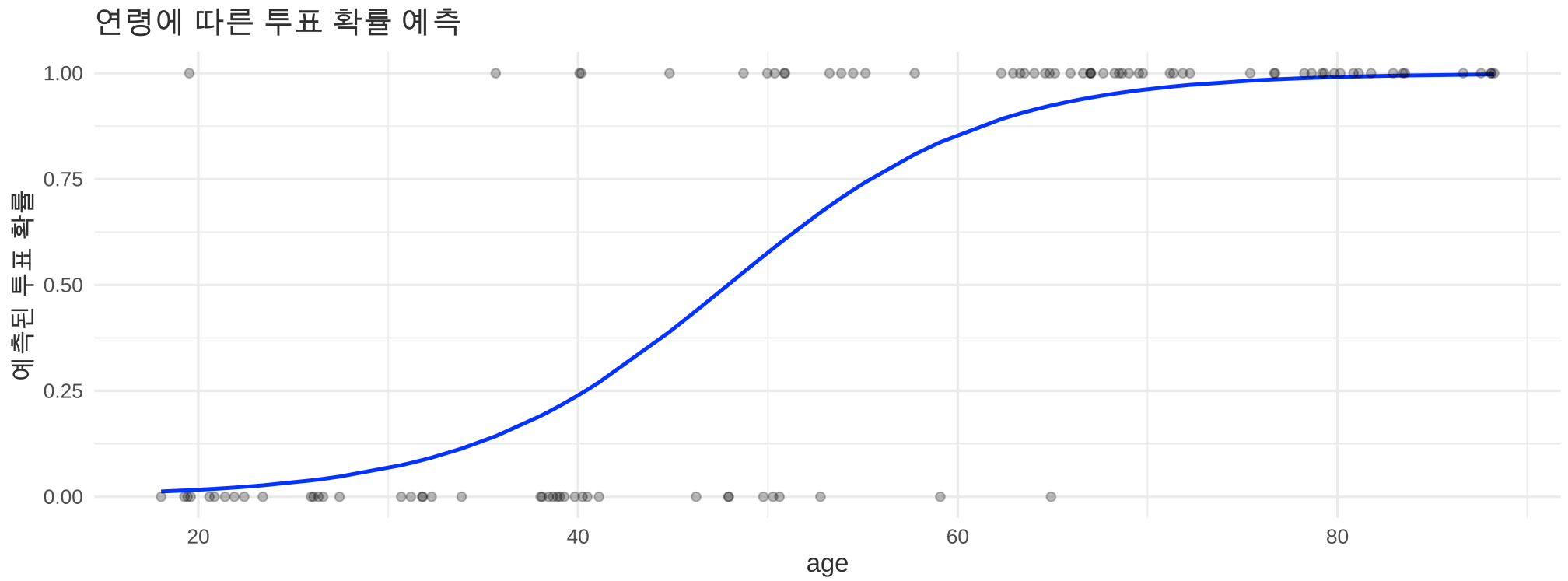
예측 확률(Predicted Probabilities)

```
# 데이터에 예측 확률 추가
data_aug <- augment(logit_model, type.predict = "response")

ggplot(data_aug, aes(x = age, y = .fitted)) +
  geom_line(color = "blue", size = 1) +
  geom_point(aes(y = vote), alpha = 0.3) +
  labs(y = "예측된 투표 확률", title = "연령에 따른 투표 확률 예측")
```

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

예측 확률(Predicted Probabilities)



비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

한계 효과(Marginal Effects)

비선형 모형에서는 X 가 변할 때 Y 가 변하는 정도(기울기)가 X 의 위치에 따라 다름.

- S자 곡선을 보면, 확률 0.5 부근에서는 기울기가 가파르고, 양 끝단(0 또는 1)에서는 완만함.

따라서 **평균적인 효과**를 요약해서 보여주는 것이 좋음.

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

한계 효과(Marginal Effects)

MEM(Marginal Effect at the Mean)

설명변수들의 평균값 위치에서의 기울기

AME(Average Marginal Effect)

모든 관측치에서 각각 기울기를 구한 뒤 평균 낸 값(추천)

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

한계 효과(Marginal Effects)

```
# {marginseffects} 패키지 이용하여 AME (Average Marginal Effect) 계산
library(marginaleffects)
ame_results <- avg_slopes(logit_model)

ame_results |> summary()
```

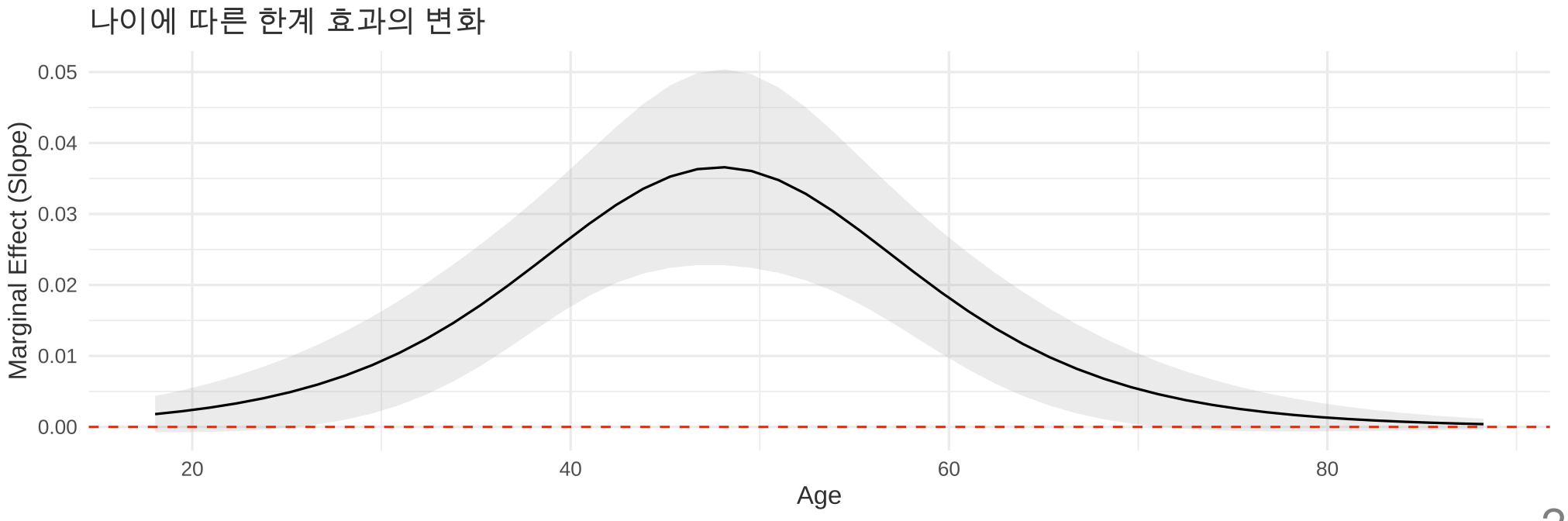
```
##      term              estimate          std.error           statistic
##  Length:1           Min.   :0.01331   Min.   :0.0001162   Min.   :114.5
##  Class :character   1st Qu.:0.01331   1st Qu.:0.0001162   1st Qu.:114.5
##  Mode  :character   Median :0.01331   Median :0.0001162   Median :114.5
##                     Mean   :0.01331   Mean   :0.0001162   Mean   :114.5
##                     3rd Qu.:0.01331   3rd Qu.:0.0001162   3rd Qu.:114.5
##                     Max.   :0.01331   Max.   :0.0001162   Max.   :114.5
##     p.value      s.value        conf.low          conf.high         predicted_lo
##  Min.   :0   Min.   :Inf   Min.   :0.01309   Min.   :0.01354   Min.   :0.9884
##  1st Qu.:0   1st Qu.:Inf   1st Qu.:0.01309   1st Qu.:0.01354   1st Qu.:0.9884
##  Median :0   Median :Inf   Median :0.01309   Median :0.01354   Median :0.9884
```

비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

한계 효과(Marginal Effects)

한계 효과(기울기)가 X 값에 따라 어떻게 변할까?



비선형 관계와 로지스틱 회귀분석

로짓 변환과 최대가능도추정(Maximum Likelihood Estimation)

한계 효과(Marginal Effects)

한계 효과(기울기)가 X 값에 따라 어떻게 변할까?

나이가 50세 근처일 때, 나이 증가가 투표 확률에 미치는 영향(기울기)이 가장 큼.

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

모형 적합도(Goodness of Fit)

로지스틱 회귀에서는 OLS의 R^2 와 정확히 같은 개념은 존재하지 않음. 대신 **유사 R^2 (Pseudo R^2)**를 사용

McFadden's R^2

$$R^2_{McFadden} = 1 - \frac{\ln L_{model}}{\ln L_{null}}$$

- L_{model} : 우리 모형의 가능도
- L_{null} : 절편만 있는(상수) 모형의 가능도
- 보통 0.2 ~ 0.4 정도면 꽤 좋은 적합도로 간주

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

모형 적합도(Goodness of Fit)

```
# {performance} 패키지를 이용한 R2 계산
library(performance)
r2_mcfadden(logit_model)
```

```
## # R2 for Generalized Linear Regression
##        R2: 0.562
##   adj. R2: 0.547
```

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

모형 적합도(Goodness of Fit)

분류표(Confusion Matrix)

- 예측 확률이 0.5 이상이면 1(투표), 미만이면 0(기권)으로 분류했을 때, 실제 값과 얼마나 일치하는가?

```
# 예측 클래스 생성 (임계값 0.5)
data_aug <- augment(logit_model, type.predict = "response") |>
  mutate(pred_class = ifelse(.fitted > 0.5, 1, 0))

# 분류표 생성
conf_mat <- table(Actual = data_aug$vote, Predicted = data_aug$pred_class)
conf_mat
```

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

모형 적합도(Goodness of Fit)

분류표(Confusion Matrix)

- 예측 확률이 0.5 이상이면 1(투표), 미만이면 0(기권)으로 분류했을 때, 실제 값과 얼마나 일치하는가?

```
conf_mat
```

```
##        Predicted
## Actual  0  1
##      0 33  8
##      1  5 54
```

정확도(Accuracy): 0.87

$$\frac{TP + TN}{}$$

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

ROC 곡선과 AUC

ROC(Receiver Operating Characteristic) 곡선

- 분류 임계값(Threshold)을 0에서 1까지 변화시키며 **민감도(True Positive Rate)**와 **1-특이도(False Positive Rate)**를 그린 곡선

- 왼쪽 위 모서리에 가까울수록 좋은 모형

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

ROC 곡선과 AUC

AUC(Area Under Curve)

- ROC 곡선 아래의 면적.
    - 0.5: 무작위 예측(나쁨)
    - 1.0: 완벽한 예측(좋음)
- 일반적으로 0.7 이상이면 수용 가능, 0.8 이상이면 우수

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

ROC 곡선과 AUC

```
# pROC 패키지 활용
library(pROC)
roc_obj <- roc(data$vote, fitted(logit_model))

# AUC 값 출력
auc(roc_obj)
```

```
## Area under the curve: 0.9467
```

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

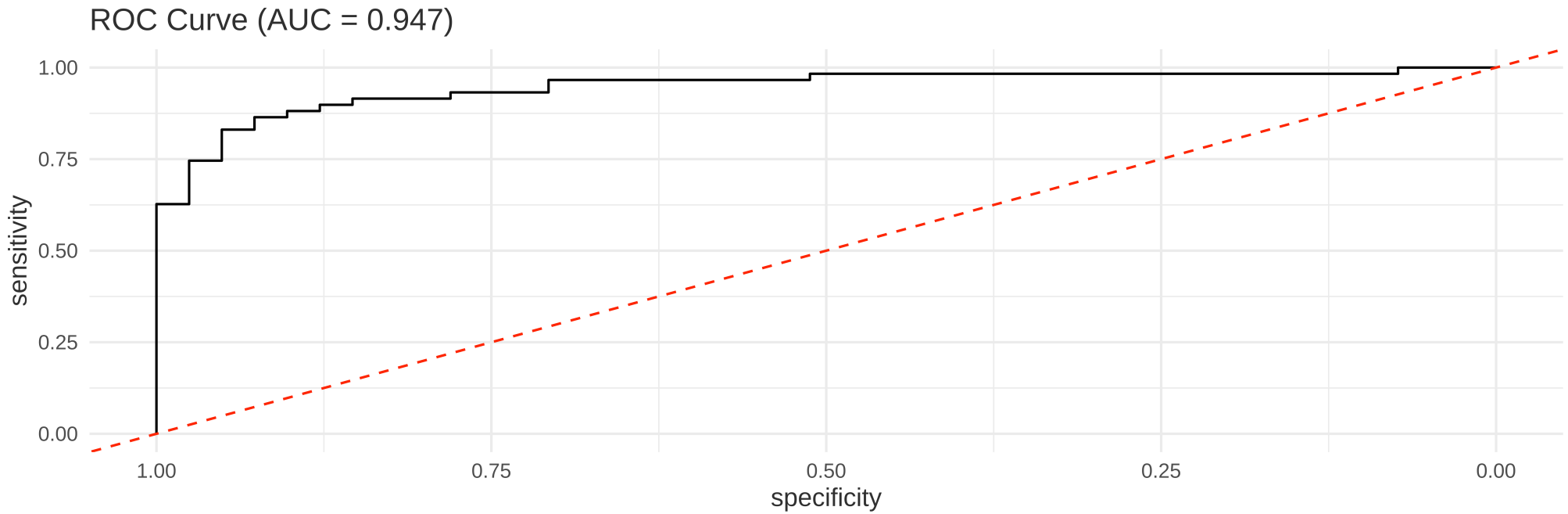
ROC 곡선과 AUC

```
# 그래프
ggroc(roc_obj) +
  geom_abline(slope = 1, intercept = 1, linetype = "dashed", color = "red") +
  labs(title = paste0("ROC Curve (AUC = ", round(auc(roc_obj), 3), ")"))
```

비선형 관계와 로지스틱 회귀분석

모형의 평가: 적합도와 분류 성능

ROC 곡선과 AUC



비선형 관계와 로지스틱 회귀분석

프로빗(Probit) 모형

로짓과 무엇이 다른가?

프로빗 모형(Probit Model)

로지스틱 회귀가 로지스틱 분포를 사용한다면, 프로빗 모형은 표준정규분포의 누적분포함수 (Φ) 를 사용

$$P(Y = 1) = \Phi(\beta_0 + \beta_1 X)$$

로짓 vs 프로빗

- 결과의 유사성: 실질적으로 두 모형의 결과(예측 확률, 한계 효과)는 거의 차이가 없음.
- 계수의 크기: 로짓 계수 $\approx 1.6 \sim 1.8 \times$ 프로빗 계수(분산 차이 때문)

비선형 관계와 로지스틱 회귀분석

프로빗(Probit) 모형

로짓과 무엇이 다른가?

```
# 프로빗 모형 적합
probit_model <- glm(vote ~ age, data = data, family = binomial(link = "probit"))
```

비선형 관계와 로지스틱 회귀분석

프로빗(Probit) 모형

로짓과 무엇이 다른가?

```
library(modelsummary)
# 결과 비교 (modelsummary 패키지)
texreg::screenreg(list("Logit" = logit_model, "Probit" = probit_model), single.row = T)
```

```
##
## ================================================
##                   Logit                Probit
## ------------------------------------------------
## (Intercept)      -7.01 (1.40) ***   -3.79 (0.68) ***
## age               0.15 (0.03) ***    0.08 (0.01) ***
## ------------------------------------------------
## AIC              63.30              63.87
## BIC              68.51              69.08
## Log Likelihood  -29.65             -29.93
## Deviance         59.30              59.87
## Num. obs.       100                100
## ================================================
```

비선형 관계와 로지스틱 회귀분석

프로빗(Probit) 모형

로짓과 무엇이 다른가?

로짓 vs 프로빗 예측 확률 비교

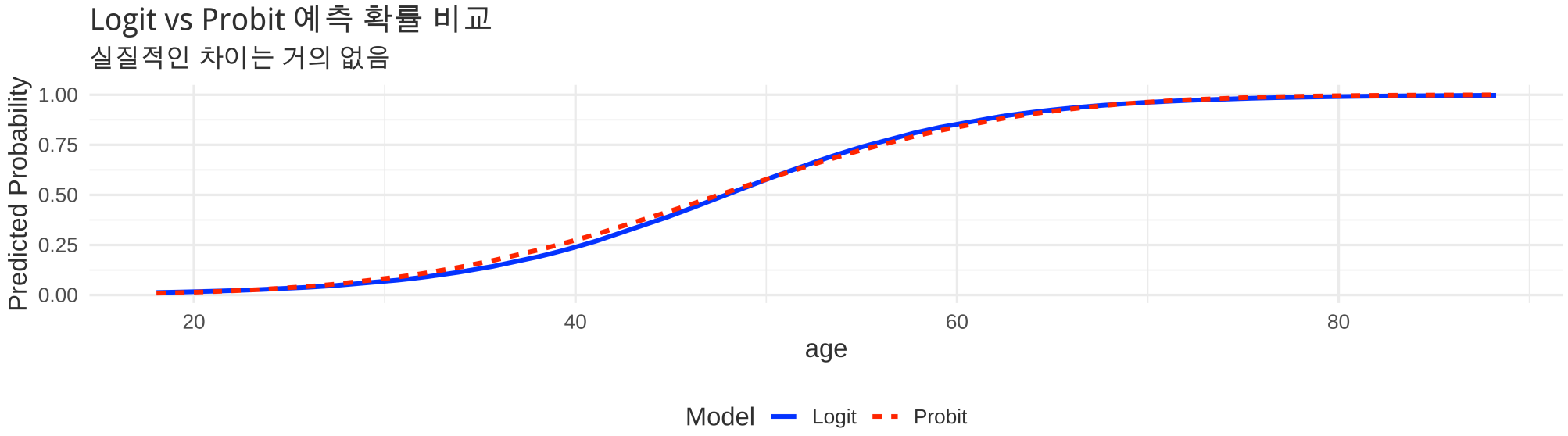
```
# marginaleffects 패키지를 이용해 두 모델의 예측값 계산
pred_logit <- predictions(logit_model) %>% mutate(Model = "Logit")
pred_probit <- predictions(probit_model) %>% mutate(Model = "Probit")

bind_rows(pred_logit, pred_probit) %>%
  ggplot(aes(x=age, y=estimate, color=Model, linetype=Model)) +
  geom_line(size=1.2) +
  labs(y="Predicted Probability", title="Logit vs Probit 예측 확률 비교",
       subtitle = "실질적인 차이는 거의 없음") +
  scale_color_manual(values=c("blue", "red")) +
  theme_minimal(base_size = 14) +
  theme(legend.position = "bottom")
```

비선형 관계와 로지스틱 회귀분석

프로빗(Probit) 모형

로짓과 무엇이 다른가?

로짓 vs 프로빗 예측 확률 비교



비선형 관계와 로지스틱 회귀분석

나가며

1. LPM의 한계: 종속변수가 0/1일 때 OLS는 예측 범위 위반, 비정규성, 이분산성 문제
2. 로지스틱 회귀: S자 곡선을 통해 확률을 [0, 1] 범위로 제한하며, MLE를 통해 추정
3. 해석
    - 계수는 Log-Odds 변화량 $\rightarrow$ 부호와 유의성 위주로 해석
    - **승산비(OR)**와 **한계효과(AME)**를 통해 구체적인 크기를 해석하는 것이 좋음.
    - `{marginaleffects}` 패키지는 AME 계산 및 시각화의 표준
    - 평가: Pseudo R^2보다는 분류표, ROC/AUC 등이 더 중요한 척도로 사용

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