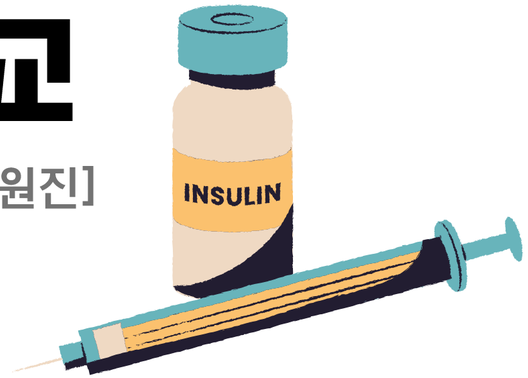
당뇨약 변경 여부에 따른 재입원과 입원일수 비교

3조 [김한얼, 서지현, 안준현, 최원진]



목 차

1

데이터 및 가설 소개

- 데이터 소개
- 변수 설명
- 가설 소개

데이터 소개

	Diabetes 130 US hospitals for years 1999-2008
	미국 내 130개 병원 및 의료 네트워크의 10년간 진행된 환자 및 병원 치료 결과를 나타내는 임상 진료 데이터

출처: Kaggle

정수형 (Int)

문자형 (Chr)

주요 변수

변수 설명

환자 기본 정보

- patient_nbr
- encounter_id
- race
- gender
- age
- weight
- medical_specialty
- payer_code

입원 및 치료정보

- admission_type_id
- admission_source_id
- discharge_disposition_id
- time_in_hospital
- num_lab_procedures
- num_procedures
- num_medications
- number_diagnoses
- number_outpatient
- number_emergency
- number_inpatient
- diag_1, diag_2, diag_3
- max_glu_serum
- A1Cresult

약물 및 결과 정보

- diabetesMed: 당뇨약 복용 여부
- change: 치료 변화 여부 (Yes: 약 변경 or 추가, No: 그대로)
- metformin, insulin, ..., glipizide (22개의 당뇨약 사용 상태)
- readmitted: 재입원 여부 ("<30", ">30", "NO")

가설 소개

1번 가설

치료 변화 여부(약 변경 or 추가)는 **재입원**에 유의한 영향을 미친다.

처리(strata): change
목표변수(y): readmitted
모델(model): 로지스틱 회귀 모형

2번 가설

치료 변화 여부(약 변경 or 추가)는 **입원 기간**에 유의한 영향을 미친다.

처리(strata): change
목표변수(y): time_in_hospital
모델(model): 선형 회귀 모형

2

전처리 & 모델링

- 결측치, 불필요한 변수 제거
- 성향점수 산출 및 매칭
- 회귀모형 적합

결측치 처리

```r
#### 데이터 로드 및 기본 결측 처리 ####
df <- read.csv("diabetic_data.csv")
sum(duplicated(df$patient_nbr))
# '?' → NA로 변환
df[df == "?"] <- NA

#### 결측률 높은 변수 제거 ####
df <- df %>% select(-c(weight, payer_code, medical_specialty))

#### 성별 이상치 제거 & 식별자 제거 ####
df <- df[df$gender != "Unknown/Invalid", ]
df <- df %>% select(-encounter_id)

#### 진단코드 보조 결측 처리 (Missing이라는 범주로 유지) ####
df$diag_2[is.na(df$diag_2)] <- "Missing"
df$diag_3[is.na(df$diag_3)] <- "Missing"

#### 주진단 결측 행 제거 ####
df <- df[!is.na(df$diag_1), ]
```

- 결측률이 50% 이상인 3개의 변수 제거
- diag_2, 3의 결측치를 MIsising이라는 범주로 대체 (보조진단 없이 주진단 (diag_1)만 존재하는 행)

중복 제거 및 변수 이진화

```r
#### 환자 중복 제거 (환자당 1건 유지) ####
df <- df %>%
  arrange(patient_nbr, desc(time_in_hospital)) %>%  # 입원 기간 긴 순서로 정렬
  group_by(patient_nbr) %>%
  slice(1) %>%   # 각 환자당 가장 긴 입원 기록 하나만 선택
  ungroup()

#### 처치 및 결과 변수 이진화 ####
df$change <- ifelse(df$change == "Ch", 1, 0)
df$readmit_bin <- ifelse(df$readmitted == "NO", 1, 0)
drug_vars <- c("insulin", "glyburide", "glipizide", "metformin")

# 이진 변수 생성 (No → 0, 나머지 → 1)
for (drug in drug_vars) {
  df[[paste0(drug, "_bin")]] <- ifelse(df[[drug]] == "No", 0, 1)
}

# A1C: 이상이면 1, 아니면 0
df$A1C_high <- ifelse(df$A1Cresult %in% c(">7", ">8"), 1, 0)

# 포도당: 이상이면 1, 아니면 0 (변수명이 'glucose'일 경우)
df$glucose_high <- ifelse(df$max_glu_serum %in% c(">200", ">300"), 1, 0)
```

- 하나의 환자가 여러 번 입원한 기록이 존재

→ 환자 당 한 건의 입원만 유지 (분석의 일관성)

- 분석의 편의를 위해 변수 이진화

새로운 변수 생성

```r
#### 당뇨병 보유 여부 변수 생성 (주 + 보조 진단코드 포함) ####
df$has_diabetes <- ifelse(grepl("^250", df$diag_1) |
                          grepl("^250", df$diag_2) |
                          grepl("^250", df$diag_3), 1, 0)

df$num_diagnoses <- rowSums(df[, c("diag_1", "diag_2", "diag_3")] != "Missing")



#### 변수 타입 정리 ####
df$has_diabetes <- as.numeric(df$has_diabetes)
df$A1C_high <- as.numeric(df$A1C_high)
df$glucose_high <- as.numeric(df$glucose_high)
df$insulin_bin <- as.numeric(df$insulin_bin)
df$metformin_bin <- as.numeric(df$metformin_bin)
df$glyburide_bin <- as.numeric(df$glyburide_bin)
df$glipizide_bin <- as.numeric(df$glipizide_bin)
# 처치 변수는 factor로 (처치군 vs 비교군 구분을 위함)
df$change <- as.factor(df$change)

# 인구통계 변수는 범주형 factor
df$age <- as.factor(df$age)
df$gender <- as.factor(df$gender)
df$race <- as.factor(df$race)

df <- df %>% select(-c(diag_1, diag_2, diag_3))
```

- 당뇨병 보유 여부: 주진단 또는 보조 진단에 ICD-9코드 상위 3자리에 250이 포함된 경우를 식별
- 진단 질환 개수 : 한 입원에서 동시에 진단 된 질환의 수 파악

성향점수(propensity score)

```r
#### 성향점수 산출 ####
df_model <- na.omit(df)

ps_model <- glm(change ~ age + gender + race +
                  A1C_high + glucose_high +
                  has_diabetes + num_diagnoses +
                  insulin_bin + metformin_bin + glipizide_bin + glyburide_bin +
                  num_medications + number_inpatient + admission_type_id,
                data = df,
                family = binomial())


df_model$pscore <- predict(ps_model, type = "response")
```

로지스틱 회귀 모형으로 성향 점수(각 환자가 처리군에 속할 확률) 계산

성향점수 매칭

```r
####  매칭 ####
library(MatchIt)

match_model <- matchit(change ~ age + gender + race +
                         A1C_high + glucose_high +
                         has_diabetes + num_diagnoses +
                         insulin_bin + metformin_bin + glipizide_bin + glyburide_bin +
                         num_medications + number_inpatient + admission_type_id,
                       data = df_model,
                       caliper = sd(df_model$pscore)*0.25,
                       method = "nearest",    # 최근접 매칭
                       ratio = 1)             # 1:1 매칭

summary(match_model)  # 매칭 전후 공변량 균형 확인 (SMD 확인)

matched_df <- match.data(match_model)
```

	Treated	Control
Mached	14,687	14,687
Unmatched	22,985	17,229

14,687 쌍이 매칭

[두 집단의 공변량 분포를 비슷하게 맞춘 환자들]

매칭 전

Summary of Balance for All Data:

	Means Treated	Means Control	Std. Mean Diff.
distance	0.7362	0.2235	1.9837
age[0-10)	0.0010	0.0032	-0.0669
age[10-20)	0.0069	0.0080	-0.0133
age[20-30)	0.0165	0.0148	0.0136
age[30-40)	0.0372	0.0376	-0.0021
age[40-50)	0.1006	0.0905	0.0338
age[50-60)	0.1803	0.1688	0.0299
age[60-70)	0.2349	0.2122	0.0535
age[70-80)	0.2516	0.2568	-0.0120
age[80-90)	0.1487	0.1766	-0.0784
age[90-100)	0.0222	0.0316	-0.0633
genderFemale	0.5215	0.5419	-0.0409
genderMale	0.4785	0.4581	0.0409
raceAfricanAmerican	0.1876	0.1835	0.0106
raceAsian	0.0065	0.0078	-0.0165
raceCaucasian	0.7642	0.7730	-0.0209
raceHispanic	0.0228	0.0207	0.0141
raceOther	0.0189	0.0150	0.0288
A1C_high	0.1687	0.0923	0.2038
glucose_high	0.0305	0.0216	0.0523
has_diabetes	0.4083	0.3460	0.1267
num_diagnoses	2.9818	2.9793	0.0159
insulin_bin	0.8002	0.2894	1.2777
metformin_bin	0.3616	0.0797	0.5867
glipizide_bin	0.2032	0.0626	0.3494
glyburide_bin	0.1743	0.0527	0.3207
num_medications	18.7347	14.4193	0.4676
number_inpatient	0.3528	0.3041	0.0564
admission_type_id	2.0776	2.0942	-0.0114

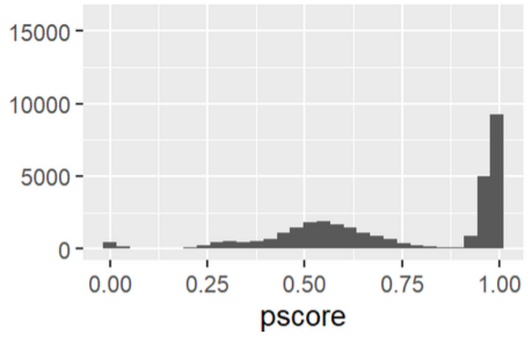
매칭 후

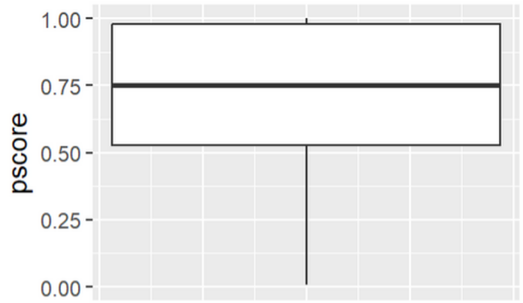
Summary of Balance for Matched Data:

	Means Treated	Means Control	Std. Mean Diff.
distance	0.4965	0.4747	0.0844
age[0-10)	0.0022	0.0017	0.0169
age[10-20)	0.0127	0.0138	-0.0132
age[20-30)	0.0242	0.0203	0.0310
age[30-40)	0.0429	0.0403	0.0137
age[40-50)	0.0985	0.0915	0.0231
age[50-60)	0.1676	0.1702	-0.0067
age[60-70)	0.2266	0.2219	0.0111
age[70-80)	0.2465	0.2523	-0.0132
age[80-90)	0.1521	0.1631	-0.0310
age[90-100)	0.0267	0.0250	0.0115
genderFemale	0.5428	0.5412	0.0033
genderMale	0.4572	0.4588	-0.0033
raceAfricanAmerican	0.1975	0.1931	0.0113
raceAsian	0.0070	0.0065	0.0060
raceCaucasian	0.7560	0.7658	-0.0231
raceHispanic	0.0219	0.0195	0.0155
raceOther	0.0176	0.0150	0.0190
A1C_high	0.1502	0.1290	0.0567
glucose_high	0.0321	0.0263	0.0336
has_diabetes	0.4098	0.3955	0.0291
num_diagnoses	2.9771	2.9771	0.0004
insulin_bin	0.7577	0.7178	0.0998
metformin_bin	0.0987	0.1232	-0.0509
glipizide_bin	0.0566	0.0638	-0.0179
glyburide_bin	0.0479	0.0546	-0.0176
num_medications	17.1311	16.2621	0.0942
number_inpatient	0.4068	0.3599	0.0543
admission_type_id	2.0742	2.0463	0.0190

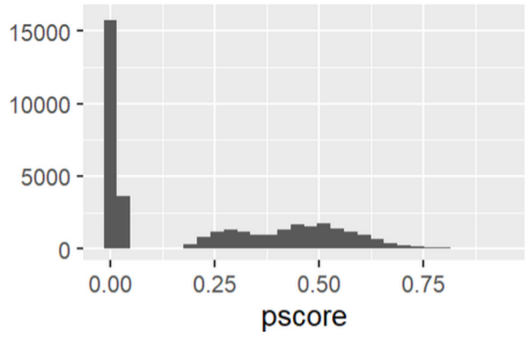
매칭 전

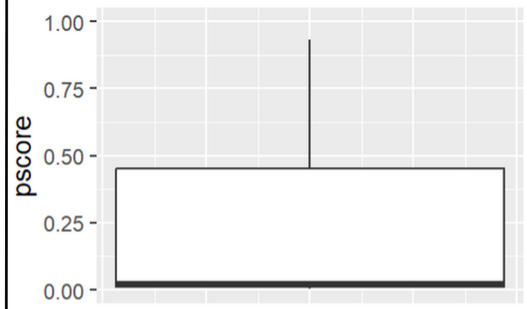
치료군





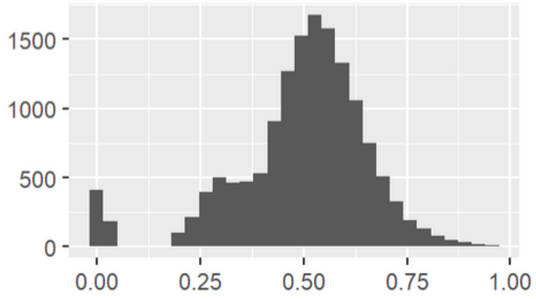
대조군

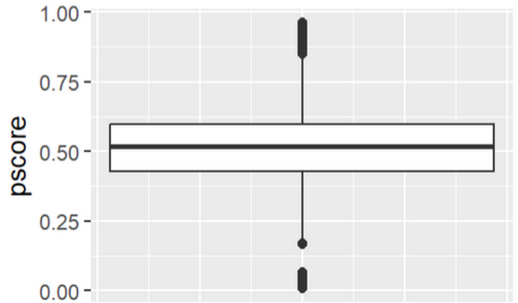




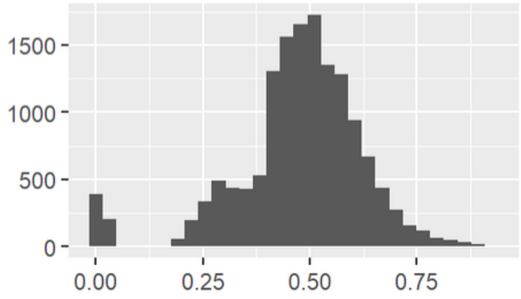
매칭 후

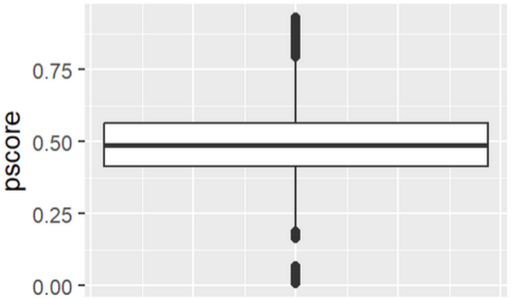
치료군





대조군





3

결과 분석

- 회귀모형 적합
- 오즈비 해석

로지스틱 회귀모형 적합

가설 1.

치료 변화 여부(약 변경 or 추가)는 **재입원**에 유의한 영향을 미친다.

```
# 로지스틱 회귀모형 적합
fit_logistic = glm(readmit_bin ~ change, data = matched_df, family = binomial())
summary(fit_logistic)
```

```
> summary(fit_logistic)

Call:
glm(formula = readmit_bin ~ change, family = binomial(), data = matched
_df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.63628    0.01735  36.683  < 2e-16 ***
change1     -0.14238    0.02429  -5.861 4.61e-09 ***
```

readmit_bin: 재입원 여부
(재입원: 1, else: 0)

change: 치료 변경 여부
(약 변경 or 추가: 1, else: 0)

적합된 회귀식:
y = 0.64 - 0.14 * change

해석:
치료법을 변경하면 재입원할 위험(오즈)이 감소한다.
따라서, **H0을 기각한다.**

오즈 해석

```r
 ### 오즈비 ###
library(epitools)


table_change_readmit <- table(matched_df$change, matched_df$readmit_bin)
print(table_change_readmit)
odds_result <- epitab(table_change_readmit,
                      method = "oddsratio",
                      rev = c("both"))
print(odds_result)
```

```r
# 오즈
OR = exp(-0.14238)
OR # 0.8672916
```

	재입원	재입원 x
약물 변경	5,083	9,604
약물 유지	5,566	3,121

치료법을 변경하지 않은 환자는 변경한 환자보다 재입원의 오즈가 약 13% 낮다.

맥니마 검정(McNemar's test)

```
## McNemar's test ##
table_change_readit = table(matched_df$change, matched_df$readmit_bin)
mcnemar.test(table_change_readit, correct = F)
```

```
> mcnemar.test(table_change_readit, correct = F)

        McNemar's Chi-squared test

data:  table_change_readit
McNemar's chi-squared = 1074.8, df = 1, p-value < 2.2e-16
```

	재입원	재입원 x
약물 변경	5,083	9,604
약물 유지	5,566	3,121

맥니마 검정 결과, p-value가 0에 근사한다.
따라서, 치료법 변경 유무에 따라
재입원 유무에 유의미한 차이가 있다고 할 수 있다.

선형회귀모형 적합

가설 2.

치료 변화 여부(약 변경 or 추가)는 **입원 기간**에 유의한 영향을 미친다.

```
# 선형회귀 모형 적합

fit_linear = lm(time_in_hospital ~ change, data = matched_df)
summary(fit_linear)
```

```
Call:
lm(formula = time_in_hospital ~ change, data = matched_df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0449 -2.7187 -0.7187  1.9551  9.2813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.71866    0.02646 178.326   <2e-16 ***
change1      0.32627    0.03742   8.719   <2e-16 ***
```

time_in_hospital : 입원 기간

change: 치료 변경 여부
(약 변경 or 추가: 1, else: 0)

적합된 회귀식:
y = 4.72 + 0.33 * change

해석:
치료법을 변경하면 입원 기간이
약 0.33일 증가한다.
따라서, **H0을 기각한다.**

eta squared

```
## eta_squared ##
library(lsr)
etaSquared(aov(time_in_hospital ~ change, matched_df))
```

```
> etaSquared(aov(time_in_hospital ~ change, matched_df))
           eta.sq eta.sq.part
change 0.0025815   0.0025815
```

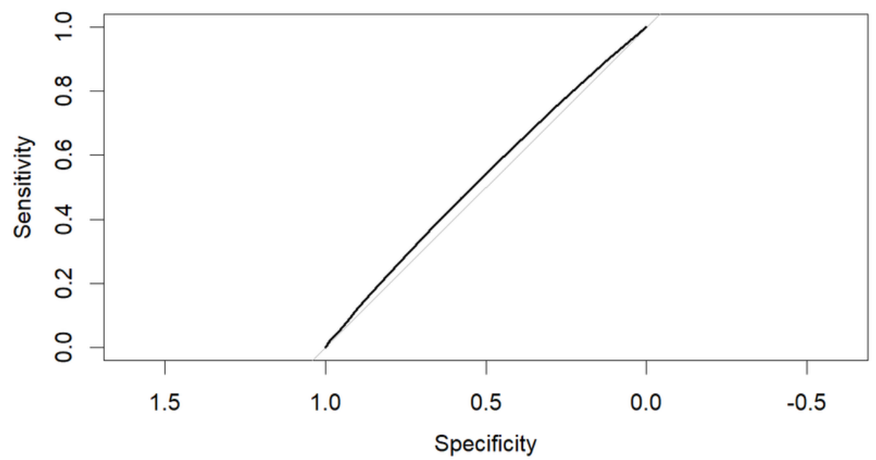
eta square = 0.003 < 0.01

→ 치료법 변경은 입원 기간에 무시해도 될 수준의 매우 작은 영향을 미친다.

ROC 곡선, AUC

```
## ROC 곡선 ##
library(pROC)
roc_obj = roc(matched_df$change, matched_df$time_in_hospital)
auc(roc_obj)
plot(roc_obj)
```

```
> auc(roc_obj)
Area under the curve: 0.5291
```



단 AUC 값이 약 0.53이고, ROC 곡선의 기울기가 1에 가깝다.
→ 분류 성능이 무작위 추측과 거의 동일하다고 할 수 있으므로 치료법 변경은 입원 기간에 **유의한 영향을 미치지 못 한다.**

감사합니다