데이터사이언스기초_
이혼에 관하여

빅데이터학과
20185254 신준철

프로젝트를 들어가며

우선 무엇보다 주제에 관한 고민에 모든 초점을 맞추었습니다. 독창적인 주제로 진행해 보고 싶은 욕심이 생겨 일반적이지 않은 주제를 찾았습니다. 여러 주제들이 흥미롭게 다가왔지만 누군가 할 법한 주제인 것 같은 느낌에 쉽게 주제를 정하지 못했습니다. 따라서 연령대를 높여서 생각해 보았습니다.

제가 정하고자 하는 주제는 20대의 생각과 고민이 투영된 문제의식에 결과물입니다. 저와 다른 세대의 생각과 고민이 지금의 저와 상반되고 그 세대의 문제의식에 의해 주제를 선정한다면 독창적일 수 있다고 생각했습니다. 따라서 각 세대별 생각과 고민이 무엇인지 살펴 보았고 그 중 40~50대의 주된 생각과 고민이 흥미롭게 다가왔습니다.

20대인 저는 취업과 연애가 주된 생각이자 고민입니다. 하지만 40대의 생각과 고민은 이직 및 은퇴, 이혼과 관련된 것이었습니다.
저는 이중 이혼과 관련된 주제가 흥미롭게 다가 왔고 그것을 주제로 택하게 되었습니다

이혼이란 것이 사회적인 흠은 결코 아니지만 평생의 사랑의 서약을 마친 '오직 두 사람'이 이혼을 통해 다른 길을 가게 되는 것에, 사랑이란 것의 끝이 있는 것인가란 슬픈 의혹을 가지게 되었습니다. 하지만 어쩌면 이혼은 '오직 두 사람'이 아닌 여러 요소들에 의한 사건일 뿐이고 가스 벨브를 잠그듯이 미연에 방지할 수 있는 일이라면 사랑이란 낭만을 더욱 소중히 간직할 수 있게 되지 않을까란 희망을 가지고 이 프로젝트를 시작하게 되었습니다.

프로젝트 초록

본 프로젝트는 이혼 여부를 종속 변수로 사용하여 데이터 분석을 수행하였습니다. 데이터를 획득하고 정제한 후, 변수들 간의 상관계수를 계산하여 상관관계를 파악하고, 이를 시각화하여 분석한 결과를 도출하였습니다. 또한, 가설 검정을 통해 변수들 간의 관계를 통계적으로 검증하고, 선형회귀 모델을 활용하여 이혼 여부를 예측하는 모델을 학습시켰습니다.

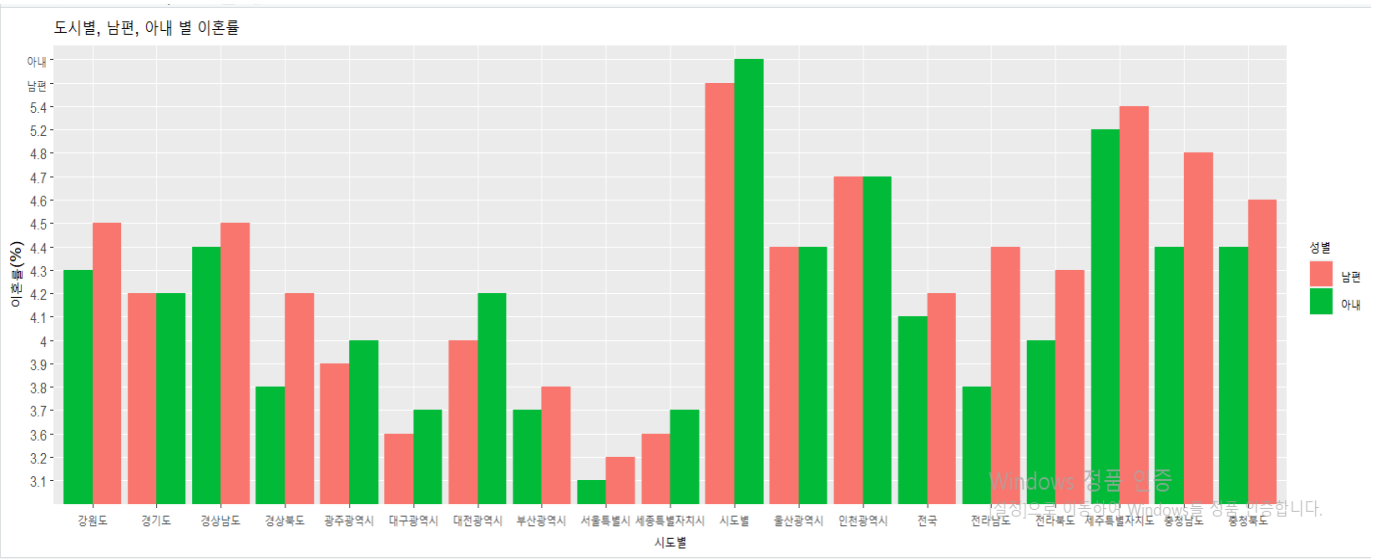
본 프로젝트는 시도별, 연령별, 직업별, 혼인 기간별 이혼율에 대해 더 깊이 탐구할 예정

입니다. 이를 위해 시각화 기법을 활용하여 데이터를 살펴보고, lm 함수를 활용하여 선형 회귀 모델을 구축하였습니다. 이를 통해 이혼율과 관련된 다양한 요인들 사이의 관계를 파악하고, 이를 통해 예측 모델을 개발하는 데 기여하였습니다.

본 프로젝트는 이혼에 영향을 미치는 요인들을 이해하고, 이혼 여부를 예측하는 데 도움이 되는 통계적인 분석과 모델링을 수행함으로써, 사회 현상을 이해하고 개인 및 정책 수립에 기여할 수 있는 정보를 제공하고자 합니다.

데이터 취득, 정제 ,가공 시각화

다음은 한국의 시도별/성별 이혼률 통계입니다.



통계청,「인구동향조사」, 2022, 2023.05.01, 시도/일반이혼율
https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B85031&conn_path=I2

```
# 데이터 불러오기
df <- read_excel("시도_일반이혼율_20230501170138.xlsx", sheet = "데이터")
df<-as_tibble(df,locale=locale("ko,encoding=EUC_KR"),na=".")
```

```
# 데이터 프레임 재구성
df_new <- df %>%
  pivot_longer(cols = -시도별, names_to = "성별", values_to = "이혼률") %>%
  mutate(성별 = if_else(성별 == "2022...2", "남편", "아내"))

# 그래프 그리기
ggplot(df_new, aes(x=시도별, y=이혼률, fill=성별)) +
  geom_bar(stat="identity", position=position_dodge()) +
  scale_fill_manual(values=c("#f8766d", "#00ba38")) +
  ggtitle("도시별, 남편, 아내 별 이혼률") +
  labs(x="시도별", y="이혼률(%)")
```

먼저 데이터 취득과 정제입니다.

tidyverse와 readxl 패키지를 설치하고 로드하여 데이터를 불러왔습니다.
read_excel 함수를 사용하여 "시도_일반이혼율_20230501170138.xlsx" 파일의 "데이터" 시트에서 데이터를 불러왔습니다.
as_tibble 함수를 사용하여 데이터를 tibble 형식으로 변환하였습니다.
데이터의 인코딩은 "EUC_KR"로 설정하였고, 결측치는 "."으로 처리하였습니다.

다음으로 데이터 가공 과정을 수행하였습니다.

데이터 프레임인 df를 시도별을 기준으로 "남편"과 "아내"의 이혼률을 구분할 수 있도록 재구성하였습니다.

pivot_longer 함수를 사용하여 "시도별" 열을 고정하고, "남편"과 "아내"의 이혼률을 "성별" 열로 변환하였습니다.
마지막으로 데이터를 시각화 하였습니다.

ggplot2 패키지를 사용하여 데이터를 시각화하였습니다.
ggplot 함수를 사용하여 기본적인 그래프 객체를 생성하였습니다.
geom_bar 함수를 사용하여 막대 그래프를 그렸습니다. stat = "identity"를 설정하여 이혼률 값을 그대로 사용하였고, position = position_dodge()를 설정하여 막대를 나란히 그렸습니다.

scale_fill_manual 함수를 사용하여 "남편"과 "아내"에 대해 원하는 색상을 설정하였습니다.
ggtitle 함수와 labs 함수를 사용하여 그래프의 제목과 축 레이블을 설정하였습니다.

전국적인 통계로 기혼자의 4% 이상이 이혼을 하게 되고 그 양상은 도시별, 성별로 다르게 나타났습니다.

가설검정과 선형회귀, 모델링과 예측 가능성

시도별 이혼률의 가설 검정을 해보겠습니다. 아래는 코드입니다.

```
# 데이터 불러오기
df <- read_excel("시도_일반이혼율_20230501170138.xlsx", sheet = "데이터")
df<-as_tibble(df,locale=locale("ko,encoding=EUC_KR"),na=".")

# 데이터 프레임 재구성
df_new <- df %>%
  pivot_longer(cols = -시도별, names_to = "성별", values_to = "이혼률") %>%
  mutate(성별 = if_else(성별 == "2022...2", "남편", "아내"))

# 그래프 그리기
ggplot(df_new, aes(x=시도별, y=이혼률, fill=성별)) +
  geom_bar(stat="identity", position=position_dodge()) +
  scale_fill_manual(values=c("#f8766d", "#00ba38")) +
  ggtitle("도시별, 남편, 아내 별 이혼률") +
  labs(x="시도별", y="이혼률(%)")
#-----------------------------------------------아래부터 추가된 코드

print(df_new)
# 결측치와 무한대 값 제거
df_new <- na.omit(df_new)
```

```
#필요 없는 값 제거 ex 전국이란 텍스트는 숫자를 비교할 때 오류가 된다.
df_new <- df_new[5:nrow(df_new), ]
df_new <- df_new %>%
  group_by(시도별) %>%
  summarise(이혼률 = mean(as.numeric(이혼률)))
```

#귀무가설 귀무가설(H0): 시에 산다면 이혼률이 낮지 않다를 선형 회귀 분석하기 위하여 이혼률을 작은 순서에서 정렬하였다. 작은 순서에서 차례대로 적용하여 시는 더 작은 순서에 배정하고 도는 그 뒤 순서에 배정하여 실제로 시에 산다면 도에 사는 것 보다 이혼률이 높은지 낮은지 판단하기 위함이다.

```
df_new <-df_new %>%
  arrange(이혼률)

# 시로 끝나는 시도에 1부터 순서대로 적용합니다.
df_modified$시도별[str_detect(df_modified$시도별, "시$")] <- 1:8

# 도로 끝나는 시도에 11부터 순서대로 적용합니다.
df_modified$시도별[str_detect(df_modified$시도별, "도$")] <- 11:19

# 수정된 tibble 출력
df_modified
df_modified$시도별 <- as.integer(df_modified$시도별)

# lm 모델을 사용하여 가설 검정 수행
model <- lm(이혼률 ~ 시도별, data = df_modified)

# 검정 결과 출력
summary(model)
```

코드의 과정에 대해 설명하겠습니다.

우선 데이터를 전처리하였습니다.

df_new 데이터프레임에서 결측치와 무한대 값을 제거하였습니다. 이는 na.omit() 함수를 사용하여 결측치를 제거하고, df_new[5:nrow(df_new),]를 통해 필요 없는 값(ex: "전국")을

제거합니다.

그 후, 시도별 변수를 기준으로 그룹화하고 이혼률 변수의 평균을 계산하여 df_new를 업데이트하였습니다.

그 후 가설 설정을 시행하기 위해 귀무가설과 대립가설을 정하였고 유의수준은 0.05라 가정했습니다.

귀무가설(H0)은 시에 산다면 이혼률이 낮지 않다. (즉, 시와 도에 따른 이혼률의 차이가 없다.)라고 정하였고
대립가설(H1)은 시에 산다면 이혼률이 낮다. (즉, 시와 도에 따른 이혼률의 차이가 있다.)고 정하였습니다.

그 후 데이터 정렬을 수행했습니다.

df_new를 이혼률을 기준으로 작은 순서에서 큰 순서로 정렬하였습니다. 이는 시에 산다면 이혼률이 낮은 순서대로 시와 도를 배정하기 위함입니다.

그 후 변수를 할당하였습니다.

시로 끝나는 시도에는 1부터 순서대로 값을 할당합니다. 이는 str_detect() 함수를 사용하여 시도별 변수의 값이 "시"로 끝나는지 확인하고, 해당 조건에 맞는 인덱스에 1부터 8까지의 값을 할당합니다.

도로 끝나는 시도에는 11부터 순서대로 값을 할당합니다. 이는 str_detect() 함수를 사용하여 시도별 변수의 값이 "도"로 끝나는지 확인하고, 해당 조건에 맞는 인덱스에 11부터 19까지의 값을 할당합니다.

이어지는 과정으로 데이터프레임을 업데이트 하였습니다.

df_modified에 수정된 시도별 변수를 반영하여 데이터프레임을 업데이트하였습니다. 이때, 시도별 변수의 데이터 타입을 정수형으로 변환합니다.

또한 선형 회귀 분석을 하였습니다.

lm() 함수를 사용하여 이혼률과 시도별 변수 간의 선형 회귀 모델을 생성하였고 모델은 df_modified 데이터를 기반으로 구축되었습니다.

summary() 함수를 사용하여 선형 회귀 모델의 검정 결과를 출력합니다. 결과에는 회귀 계수, t-value, p-value 등이 포함됩니다.
주요한 결과로는 시도별 변수의 p-value 값이 제공되는데, 이를 통해 귀무가설을 검정합니다.

만약, 시도별 변수의 p-value 값이 유의수준(일반적으로 0.05)보다 작은 경우, 귀무가설을 기각하고 대립가설을 채택합니다. 이는 시에 산다면 이혼률이 낮다는 가설을 지지하는 결과입니다.

반대로, p-value 값이 유의수준보다 크거나 같은 경우, 귀무가설을 기각할 증거가 부족하므로 귀무가설을 채택합니다. 이는 시에 산다면 이혼률이 낮지 않다는 가설을 지지하는 결과입니다.

아래는 검정 결과 입니다.

```
> summary(model)

Call:
lm(formula = 이혼률 ~ 시도별, data = df_modified)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4304 -0.2078 -0.1045  0.1051  0.6567

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.51428    0.14823  23.708 2.65e-13 ***
시도별       0.06612    0.01277   5.176 0.000113 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3047 on 15 degrees of freedom
Multiple R-squared:  0.6411,    Adjusted R-squared:  0.6172
F-statistic: 26.79 on 1 and 15 DF,  p-value: 0.0001128
```

검정 결과를 해석하겠습니다.

먼저 회귀 계수(Coefficients)를 해석하겠습니다.

(Intercept)은 시도별이 0일 때(도에 해당할 때)의 이혼률을 나타냅니다. 추정된 회귀 계수

는 3.51428입니다. 이 값은 통계적으로 유의미한 것으로 나타나며, 시도별이 0일 때의 이혼률은 3.51428로 추정됩니다.

시도별은 시도별 변수에 대한 회귀 계수는 0.06612입니다. 이 값은 시도별 변수의 증가가 이혼률에 어떤 영향을 미치는지를 나타내는 것으로 해석됩니다.

t-value와 p-value를 살펴보겠습니다.

t value는 시도별 변수의 t-value는 5.176입니다. 이 값은 시도별 변수의 회귀 계수를 표준 오차로 나눈 것으로, 해당 변수가 통계적으로 유의미한 영향을 미치는지를 평가하는 지표입니다.

Pr(>|t|): 시도별 변수의 p-value는 0.000113입니다. 이 값은 시도별 변수가 이혼률에 유의미한 영향을 미치는지를 나타내는 것으로 해석됩니다. 여기서 p-value는 유의수준 0.05보다 작으므로 시도별 변수는 이혼률에 유의미한 영향을 미친다고 할 수 있습니다.
회귀 모델의 적합도:

Multiple R-squared: 이 모델의 다중 결정 계수는 0.6411입니다. 이 값은 시도별 변수가 종속 변수인 이혼률의 변동을 약 64.11% 설명한다는 것을 의미합니다. 다른 설명 변수가 누락되어 있을 수 있으므로 주의가 필요합니다.

Adjusted R-squared: 조정된 다중 결정 계수는 0.6172입니다. 이 값은 모델의 복잡성을 고려하여 다중 결정 계수를 조정한 값으로, 예측력을 보다 신뢰할 수 있는 지표입니다.

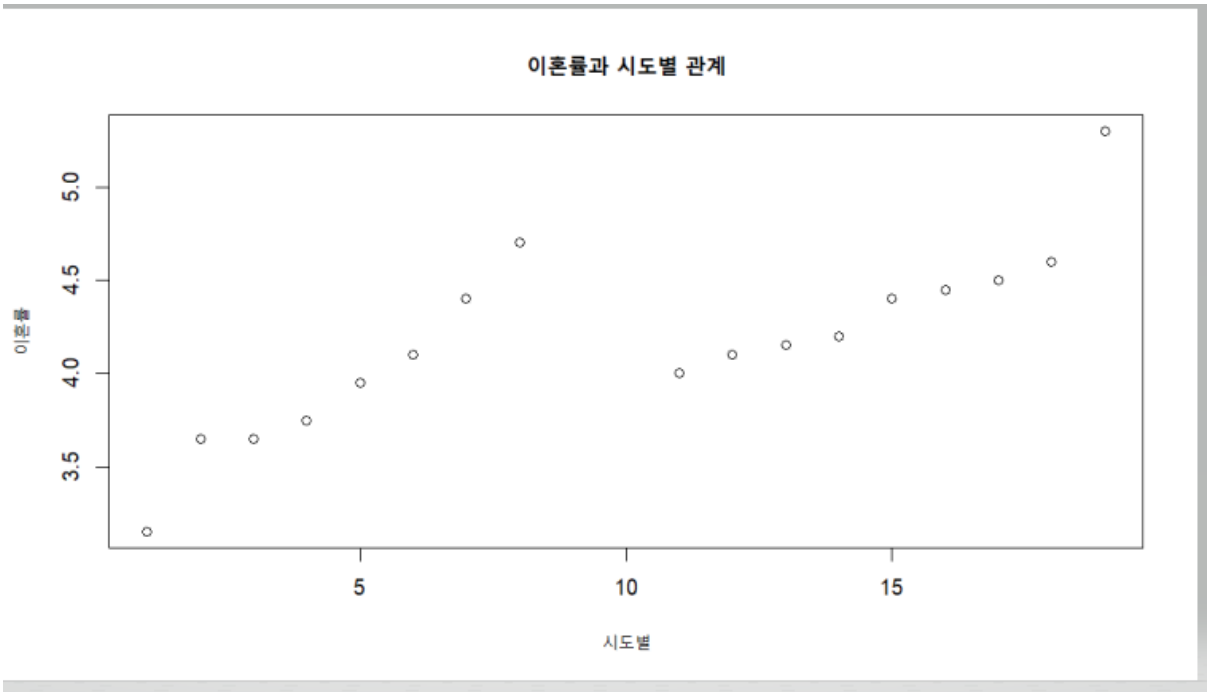
F-statistic와 p-value:
F-statistic: F-통계량은 26.79입니다. 이 값은 회귀 모델의 전체적인 유의성을 검정하는 데 사용됩니다. F-통계량이 크고 p-value가 작을수록 모델이 유의미한 것으로 판단됩니다.
p-value: F-statistic의 p-value는 0.0001128입니다. 이 값은 회귀 모델이 유의미한지를 나타내는 것으로, 유의수준 0.05보다 작으므로 회귀 모델은 유의미한 것으로 판단됩니다.
따라서, 이 결과를 종합해 보면 시도별 변수가 이혼률에 유의미한 영향을 미친다고 할 수 있으며, 모델 자체도 유의미한 예측력을 가지고 있다고 할 수 있습니다.

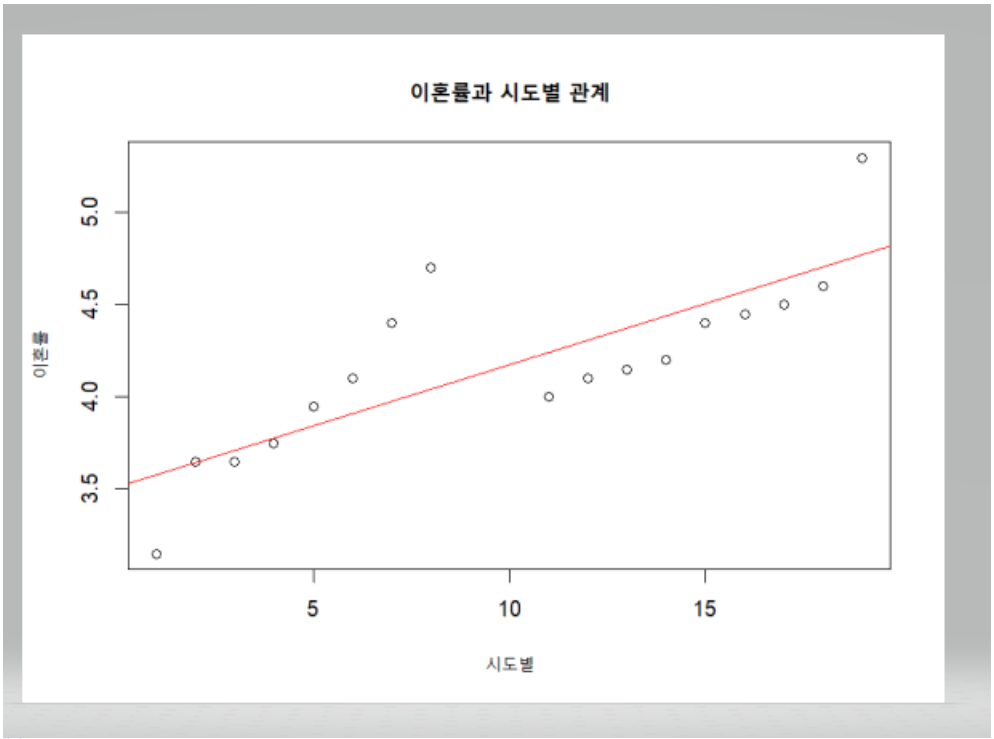
다음으로 모델링과 예측 가능성을 판별해 보겠습니다.

아래는 가능성을 판별하기 위한 코드입니다.

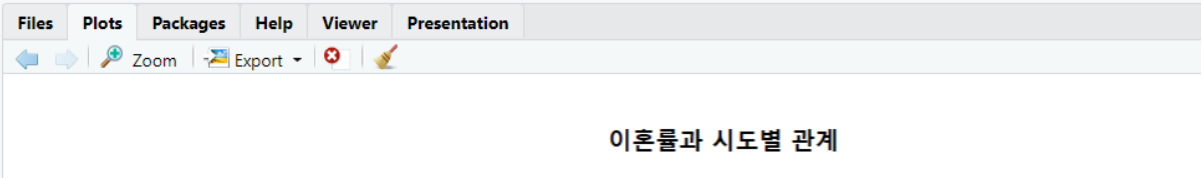
```
# 데이터 탐색
plot(df_modified$시도별, df_modified$이혼률, main = "이혼률과 시도별 관계", xlab = "시도별", ylab = "이혼률")

# 상관 관계 분석
correlation <- cor(df_modified$시도별, df_modified$이혼률)
print(correlation)

# 시각화
plot(df_modified$시도별, df_modified$이혼률, main = "이혼률과 시도별 관계", xlab = "시도별", ylab = "이혼률")
abline(lm(df_modified$이혼률 ~ df_modified$시도별), col = "red")
```

아래는 데이터 탐색 , 시각화 결과, 상관 관계 분석 결과 입니다.






```
> # 상관 관계 분석
> correlation <- cor(df_modified$시도별, df_modified$이혼률)
> print(correlation)
[1] 0.8006811
>
```



위의 시각화와 상관 관계 분석을 통해 다음과 같은 판단을 할 수 있습니다:

시각화 결과, 이혼률과 시도별 간에 어느 정도의 선형적인 관계가 있음을 확인할 수 있습니다. 선형 회귀 모델이 이 데이터에 적합하다는 시각적인 증거가 있습니다.

상관 계수 결과, 이혼률과 시도별 간의 상관 계수는 0.8006811로 양의 상관 관계가 있음을 나타냅니다. 이는 두 변수 간에 강한 선형적 관계가 있음을 시사합니다.

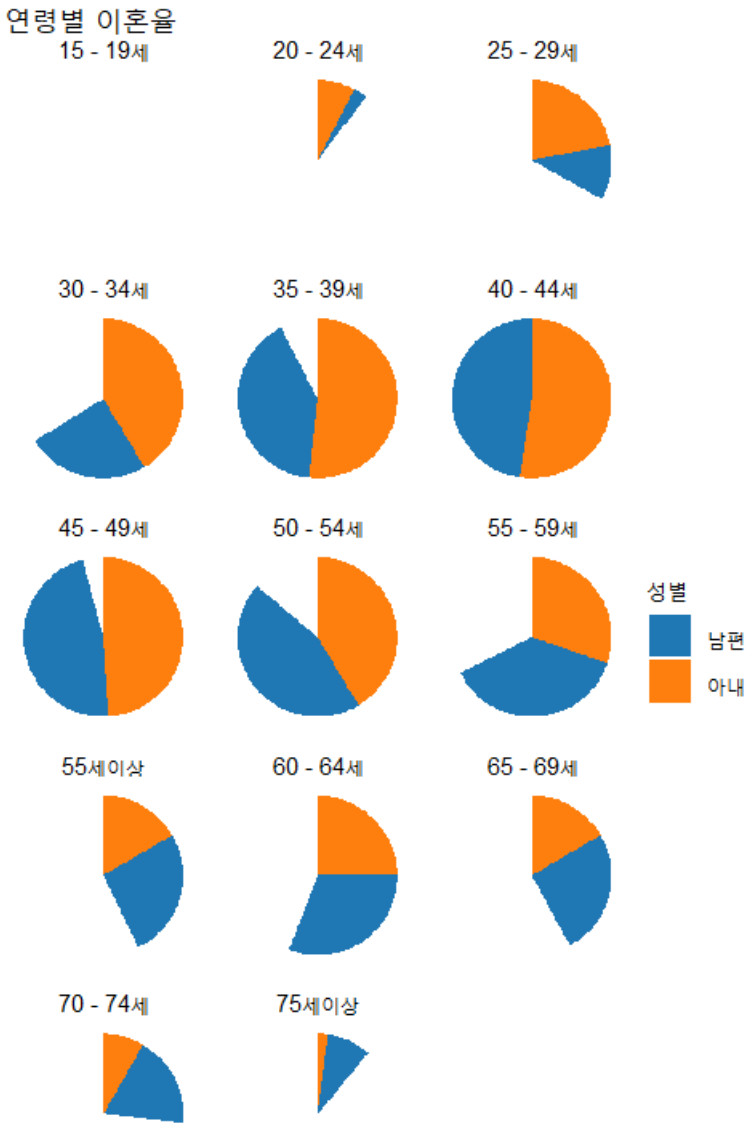
따라서, 이 데이터에서 이혼률과 시도별 간의 관계를 모델링하고 예측하는 것이 가능하며, 선형 회귀 모델이 적절한 모델링 방법일 수 있습니다.

데이터 취득, 정제 ,가공 시각화

우선 **연령별 이혼률을** 파악하기 위해 통계청,「인구동향조사」, 2022, 2023.05.02, 시도/성/연령별 이혼율 의 정보를 불러 왔습니다. 링크는 다음과 같습니다.

https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B85009&conn_path=I2

아래는 연령별 이혼률을 **시각화** 한 것입니다.



아래는 해당 **시각화를 위해 데이터를 취득, 정제, 가공**의 과정을 거친 코드입니다.

```r
# read_excel() 함수를 사용하여 엑셀 파일을 불러옵니다.
library(readxl)
data <- read_excel("시도_성_연령별_이혼율_20230502000730.xlsx")

# 데이터프레임으로 변환합니다.
df <- as.data.frame(data)
df
# 필요 없는 열을 제거합니다.
df <- df[, c(1, 2, 3, 4)]
df
# 열 이름을 변경합니다.
colnames(df) <- c("시도별", "연령별", "남편", "아내")

# 계 열을 제거합니다.
df <- df[df$시도별 != "계", ]
df
# 남편과 아내 열의 값을 천명당 건수에서 건수로 변경합니다.
df$남편 <- df$남편 * 10
df$아내 <- df$아내 * 10

# 결과를 출력합니다.
print(df)

data<-head(data,14)
data



# 나이별 이혼율을 데이터프레임으로 만들기
df_age <- data.frame(
  연령별 = data$연령별,
  이혼율 = c(data$남편, data$아내),
  성별 = rep(c("남편", "아내"), each = nrow(data))
)
```

```
# 원 그래프 그리기
ggplot(df_age, aes(x = "", y = 이혼율, fill = 성별)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  facet_wrap(~ 연령별, ncol = 3) +
  labs(title = "연령별 이혼율", x = NULL, y = NULL) +
  scale_fill_manual(values = c("#1F77B4", "#FF7F0E")) +
  theme_void()
```

주어진 R 언어 코드는 "시도_성_연령별_이혼율_20230502000730.xlsx" 파일을 데이터로 사용하여 데이터 취득, 정제, 가공 및 시각화를 수행하는 과정을 포함하고 있습니다.

데이터 취득:
read_excel 함수를 사용하여 "시도_성_연령별_이혼율_20230502000730.xlsx" 엑셀 파일을 읽어옵니다. 이 엑셀 파일에는 시도, 성별, 연령별에 따른 이혼율 데이터가 포함되어 있습니다.

데이터 정제:
불필요한 열을 제거하기 위해 df <- df[, c(1, 2, 3, 4)]를 사용하여 필요한 열만 선택합니다. 이 코드에서는 시도별, 연령별, 남편, 아내 열을 선택하여 데이터프레임 df를 생성합니다.

열 이름 변경:
colnames(df) <- c("시도별", "연령별", "남편", "아내")를 사용하여 열 이름을 한글로 변경합니다. 이는 시각화에서 더 직관적인 그래프를 생성하기 위한 작업입니다.

데이터 가공:
df <- df[df$시도별 != "계",]를 사용하여 "계"에 해당하는 행을 제거하여 시도별 데이터만 추출합니다. 이는 전체 이혼율인 "계"를 제외하고 시도별 이혼율만을 분석하기 위한 작업입니다.

천명당 건수에서 건수로 변환:
df$남편 <- df$남편 * 10와 df$아내 <- df$아내 * 10를 사용하여 남편과 아내 열의 값을 천명당 건수에서 건수로 변경합니다. 이는 시각화에서 보다 명확한 값 비교를 위한 작업

입니다.

데이터 시각화:
ggplot 함수를 사용하여 데이터를 시각화합니다. geom_bar를 사용하여 막대 그래프를 그리고, coord_polar로 극 좌표계로 변환합니다. 이를 통해 원 그래프 형태로 연령별로 남편과 아내의 이혼율을 시각화할 수 있습니다.

facet_wrap를 사용하여 연령별로 그래프를 분할합니다. 이는 연령 그룹별로 이혼율의 패턴을 비교하기 위한 작업입니다.

labs 함수를 사용하여 그래프의 제목과 축 레이블을 설정합니다. 이를 통해 그래프의 의미를 명확히 전달할 수 있습니다.

scale_fill_manual을 사용하여 성별에 따라 색상을 지정합니다. 이는 남편과 아내를 구분하기 위한 작업입니다.

theme_void를 사용하여 그래프의 배경을 투명하게 설정합니다. 이는 그래프 자체에 초점을 맞추기 위한 작업입니다.

위의 과정을 통해 데이터를 취득, 정제, 가공하여 연령별로 남편과 아내의 이혼율을 비교하는 원 그래프를 시각화할 수 있습니다. 이를 통해 연령별로 이혼율의 패턴과 남편과 아내 간의 이혼율 차이를 시각적으로 확인할 수 있습니다.

또한 20대부터 75세 이상까지 남편과 아내의 이혼률의 비가 남편이 점차 증가하는 것을 확인할 수 있습니다.

그리고 이혼율은 40-44세 까지 지속적으로 증가하다가 40-44세 이후로 지속적으로 감소하는 것을 확인할 수 있습니다.

가설검정과 선형회귀, 모델링과 예측 가능성

다음으로 연령별 이혼률을 가설 검정을 해보겠습니다. 아래는 코드입니다.

```
# read_excel() 함수를 사용하여 엑셀 파일을 불러옵니다.
library(readxl)
data <- read_excel("시도_성_연령별_이혼율_20230502000730.xlsx")

# 데이터프레임으로 변환합니다.
df <- as.data.frame(data)
df
# 필요 없는 열을 제거합니다.
df <- df[, c(1, 2, 3, 4)]
df
# 열 이름을 변경합니다.
colnames(df) <- c("시도별", "연령별", "남편", "아내")

# 계 열을 제거합니다.
df <- df[df$시도별 != "계", ]
df
# 남편과 아내 열의 값을 천명당 건수에서 건수로 변경합니다.
df$남편 <- df$남편 * 10
df$아내 <- df$아내 * 10

# 결과를 출력합니다.
print(df)

data<-head(data,14)
data


# 나이별 이혼율을 데이터프레임으로 만들기
df_age <- data.frame(
  연령별 = data$연령별,
```

```
  이혼율 = c(data$남편, data$아내),
  성별 = rep(c("남편", "아내"), each = nrow(data))
)

# 원 그래프 그리기
ggplot(df_age, aes(x = "", y = 이혼율, fill = 성별)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +
  facet_wrap(~ 연령별, ncol = 3) +
  labs(title = "연령별 이혼율", x = NULL, y = NULL) +
  scale_fill_manual(values = c("#1F77B4", "#FF7F0E")) +
  theme_void()

#------------------------------------------------아래부터 추가된 코드드

# 연령별 이혼율 계산
for (i in 1:nrow(df_age)) {
  age <- df_age$연령별[i]
  df_age$이혼율[i] <- mean((df$남편[df$연령별 == age] + df$아내[df$연령별 == age]) /
2)
}

# 나이별 값을 int 형으로 변경합니다.
df_age$연령별 <- as.integer(gsub("[^0-9]+", "", df_age$연령별))

# 연령별 값이 55와 75인 행 제거
df_age <- df_age[!(df_age$연령별 %in% c(55, 75)), ]

# 1행부터 12행까지 추출
df_age <- df_age[1:12, ]

#성별 열 삭제제
df_age <- df_age[, c("연령별", "이혼율")]

# 45세까지 이혼율 증가 가설에 대한 lm 모델링
```

model1 <- lm(이혼율 ~ 연령별, data = df_age[df_age$연령별 <= 4044,])

# 45세 이후 이혼율 감소 가설에 대한 lm 모델링
model2 <- lm(이혼율 ~ 연령별, data = df_age[df_age$연령별 > 4044,])

모델링 결과 분석
summary(model1) # 45세까지 이혼율 증가 가설에 대한 결과 분석
summary(model2) # 45세 이후 이혼율 감소 가설에 대한 결과 분석

추가된 코드에 관해 설명하겠습니다.

for 루프를 사용하여 연령별 이혼율을 계산합니다. 각 연령에 대해 남편과 아내의 이혼율을 평균 내어 해당 연령의 이혼율로 대체합니다. 이를 통해 나이별 이혼율을 계산할 수 있습니다. 따라서 df_age는 아래와 같이 변경되었습니다.
변경전

```
> df_age
       연령별 이혼율 성별
1   15 - 19세    0.0 남편
2   20 - 24세    0.4 남편
3   25 - 29세    1.6 남편
4   30 - 34세    3.6 남편
5   35 - 39세    5.9 남편
6   40 - 44세    6.9 남편
7   45 - 49세    6.8 남편
8   50 - 54세    6.5 남편
9    55세이상    3.8 남편
10 55 - 59세    5.4 남편
11 60 - 64세    4.5 남편
12 65 - 69세    3.7 남편
13 70 - 74세    2.7 남편
14  75세이상    1.3 남편
15 15 - 19세    0.0 아내
16 20 - 24세    1.1 아내
17 25 - 29세    3.2 아내
18 30 - 34세    6.0 아내
19 35 - 39세    7.5 아내
20 40 - 44세    7.6 아내
21 45 - 49세    7.1 아내
22 50 - 54세    6.0 아내
23  55세이상    2.4 아내
24 55 - 59세    4.4 아내
25 60 - 64세    3.6 아내
```

변형후

```
> df_age
       연령별       이혼율 성별
1   15 - 19세  0.08823529 남편
2   20 - 24세  7.58823529 남편
3   25 - 29세 27.52941176 남편
4   30 - 34세 52.47058824 남편
5   35 - 39세 72.50000000 남편
6   40 - 44세 76.70588235 남편
7   45 - 49세 72.91176471 남편
8   50 - 54세 62.82352941 남편
9    55세이상 29.70588235 남편
10  55 - 59세 47.91176471 남편
11  60 - 64세 38.94117647 남편
12  65 - 69세 28.64705882 남편
13  70 - 74세 18.82352941 남편
14   75세이상  7.50000000 남편
15  15 - 19세  0.08823529 아내
16  20 - 24세  7.58823529 아내
17  25 - 29세 27.52941176 아내
18  30 - 34세 52.47058824 아내
19  35 - 39세 72.50000000 아내
20  40 - 44세 76.70588235 아내
21  45 - 49세 72.91176471 아내
22  50 - 54세 62.82352941 아내
23   55세이상 29.70588235 아내
24  55 - 59세 47.91176471 아내
25  60 - 64세 38.94117647 아내
```

gsub 함수를 사용하여 연령별 값을 정수형으로 변환합니다. gsub("[^0-9]+", "", df_age$연령별)를 통해 연령별 값에서 숫자를 제외한 모든 문자를 제거하고, as.integer()를 사용하여 정수형으로 변환합니다. 이는 가설 검정을 위해 연령별 값을 정수로 사용하기 위한 작업입니다.

```
> df_age
   연령별       이혼율 성별
1    1519  0.08823529 남편
2    2024  7.58823529 남편
3    2529 27.52941176 남편
4    3034 52.47058824 남편
5    3539 72.50000000 남편
6    4044 76.70588235 남편
7    4549 72.91176471 남편
8    5054 62.82352941 남편
9      55 29.70588235 남편
10   5559 47.91176471 남편
11   6064 38.94117647 남편
12   6569 28.64705882 남편
13   7074 18.82352941 남편
14     75  7.50000000 남편
15   1519  0.08823529 아내
16   2024  7.58823529 아내
17   2529 27.52941176 아내
18   3034 52.47058824 아내
19   3539 72.50000000 아내
20   4044 76.70588235 아내
```

df_age$연령별 %in% c(55, 75)를 사용하여 연령별 값이 55 또는 75인 행을 제거합니다. 이는 55세 이상과 75세 이상의 행은 유의미하지 않기 때문에 하는 작업입니다.

```
> df_age
   연령별       이혼율 성별
1    1519  0.08823529 남편
2    2024  7.58823529 남편
3    2529 27.52941176 남편
4    3034 52.47058824 남편
5    3539 72.50000000 남편
6    4044 76.70588235 남편
7    4549 72.91176471 남편
8    5054 62.82352941 남편
10   5559 47.91176471 남편
11   6064 38.94117647 남편
12   6569 28.64705882 남편
13   7074 18.82352941 남편
15   1519  0.08823529 아내
16   2024  7.58823529 아내
17   2529 27.52941176 아내
18   3034 52.47058824 아내
19   3539 72.50000000 아내
20   4044 76.70588235 아내
21   4549 72.91176471 아내
22   5054 62.82352941 아내
24   5559 47.91176471 아내
25   6064 38.94117647 아내
26   6569 28.64705882 아내
27   7074 18.82352941 아내
```

df_age[1:12,]를 사용하여 1행부터 12행까지의 데이터를 추출합니다. 이는 13행부터 24행까지의 데이터가 이전과 같은 내용을 반복하기 때문에 1행부터 12행까지의 데이터만 사용하고자 하는 것입니다.

```
> df_age
   연령별       이혼율 성별
1    1519  0.08823529 남편
2    2024  7.58823529 남편
3    2529 27.52941176 남편
4    3034 52.47058824 남편
5    3539 72.50000000 남편
6    4044 76.70588235 남편
7    4549 72.91176471 남편
8    5054 62.82352941 남편
10   5559 47.91176471 남편
11   6064 38.94117647 남편
12   6569 28.64705882 남편
13   7074 18.82352941 남편
```

df_age[, c("연령별", "이혼율")]를 사용하여 성별 열을 삭제합니다. 이는 가설 검정을 위해 연령별과 이혼율만을 포함하는 데이터프레임을 만들기 위한 작업입니다.

```
> #성별 열 삭제제
> df_age <- df_age[, c("연령별", "이혼율")]
> df_age
   연령별       이혼율
1    1519  0.08823529
2    2024  7.58823529
3    2529 27.52941176
4    3034 52.47058824
5    3539 72.50000000
6    4044 76.70588235
7    4549 72.91176471
8    5054 62.82352941
10   5559 47.91176471
11   6064 38.94117647
12   6569 28.64705882
13   7074 18.82352941
```

가설 분석 대상에 대한 가설 검정 단계는 다음과 같습니다:

귀무가설(H0): 45세까지 이혼율은 증가하지 않는다.
대립가설(H1): 45세까지 이혼율은 증가한다.

귀무가설(H0): 45세 이후 이혼율은 감소하지 않는다.
대립가설(H1): 45세까지 이혼율은 감소한다.

유의수준을 0.05로 설정하겠습니다.

또한 선형회귀 모델을 사용하여 연령별 이혼율과 연령별 사이의 관계를 분석합니다. lm() 함수를 사용하여 모델을 생성하고, summary() 함수를 사용하여 회귀 모델의 결과를 분석합니다.

summary(model1)을 사용하여 45세까지 이혼율 증가 가설에 대한 결과를 분석합니다.

model1의 결과입니다.

```
> summary(model1)  # 45세까지 이혼율 증가 가설에 대한 결과 분석

Call:
lm(formula = 이혼율 ~ 연령별, data = df_age[df_age$연령별 <=
    4044, ])

Residuals:
     1      2      3      4      5      6
 3.662 -6.059 -3.340  4.379  7.187 -5.829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -55.37619    8.88276  -6.234 0.003373 **
연령별        0.03410    0.00305  11.180 0.000364 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.444 on 4 degrees of freedom
Multiple R-squared:  0.969,     Adjusted R-squared:  0.9612
F-statistic:   125 on 1 and 4 DF,  p-value: 0.0003644
```

결과를 해석해보겠습니다.

회귀 계수(Coefficients)는 (Intercept): -55.37619, 연령별: 0.03410 입니다.

회귀 분석 결과를 분석하겠습니다. 연령별의 회귀 계수의 p-값은 0.000364입니다. 이 값은 유의수준 0.05보다 훨씬 작으므로 통계적으로 유의미한 결과입니다.
연령별의 회귀 계수의 부호는 연령별의 회귀 계수는 양수인 0.03410입니다.

회귀 계수를 해석해보겠습니다.

(Intercept): 절편은 -55.37619입니다. 이는 연령이 0일 때의 이혼율을 나타냅니다. 이 값은 통계적으로 유의미하며, 음수인 경우에는 해석이 어렵습니다.
연령별: 회귀 계수는 0.03410입니다. 이는 연령이 1단위 증가할 때마다 이혼율이 0.03410 증가한다는 것을 의미합니다.

회귀 분석의 적합도를 알아보겠습니다.

Multiple R-squared: 0.969
Adjusted R-squared: 0.9612
F-statistic: 125, p-value: 0.0003644

이 모델은 R-squared 값이 0.969로, 이혼율의 96.9%의 변동을 연령별로 설명할 수 있습니다. Adjusted R-squared 값은 0.9612로, 모델에 포함된 변수의 수를 고려한 조정된 결정 계수입니다. F-통계량은 125이며, p-값은 0.0003644입니다. 이는 모델이 통계적으로 유의미하다는 것을 나타냅니다.

따라서, 모델1의 결과는 다음을 의미합니다:

45세까지 이혼율은 연령이 증가함에 따라 증가합니다.
모델은 이혼율의 96.9%의 변동을 연령별로 설명할 수 있으며, 이는 통계적으로 유의미합니다.

다음으로 model1의 모델링과 예측 가능성을 판별해 보겠습니다.

앞서 살펴본 바와 같이 model1은 다음과 같은 선형 회귀 모델입니다:

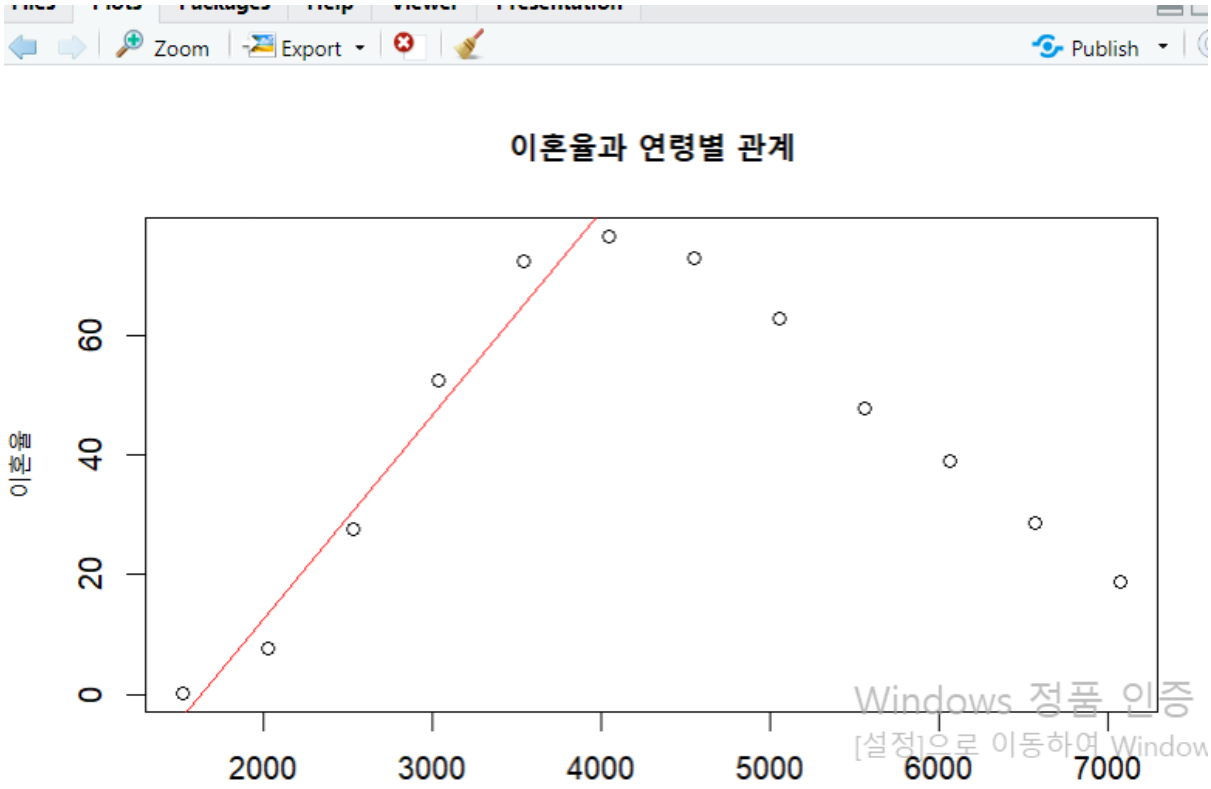
이혼율 = -55.37619 + 0.03410 * 연령별

이 모델은 연령별과 이혼율 간의 선형 관계를 나타내는 회귀식입니다. 회귀식에서 절편은 -55.37619이고, 연령별의 계수는 0.03410입니다.

아래는 가능성을 판별하기 위한 코드입니다.

# 데이터 탐색
plot(df_age$연령별, df_age$이혼율, main = "이혼율과 연령별 관계", xlab = "연령별", ylab = "이혼율")

# 모델1 시각화
abline(model1, col = "red")

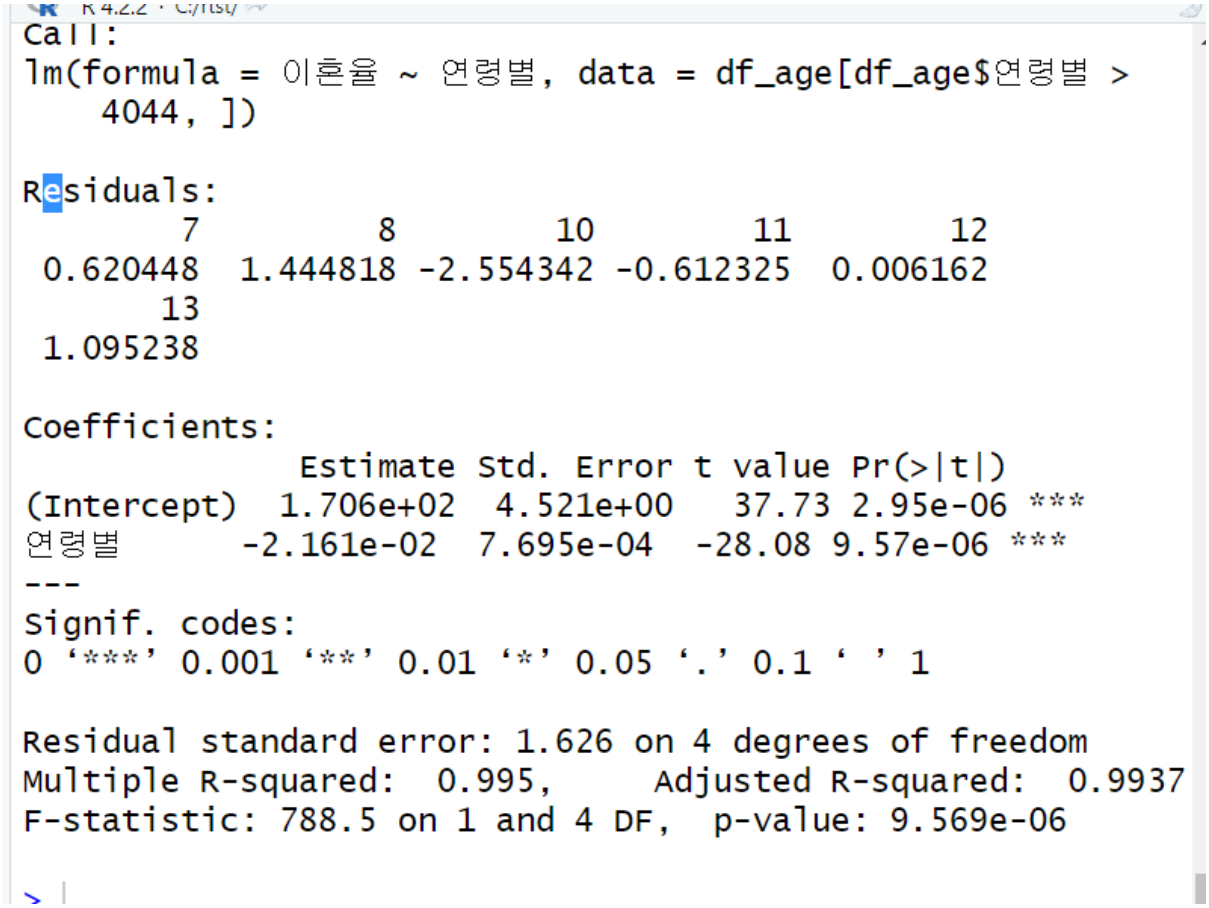


시각화 결과 이혼률과 model1(45세 이전까지의 이혼률) 간에 어느 정도의 선형적인 관계가 있음을 확인할 수 있습니다.

따라서 이 데이터에서 model1을 통해 예측하는 것이 가능하며, 선형 회귀 모델이 적절한 모델링 방법일 수 있습니다.

이어지는 과정으로 summary(model2)을 사용하여 45이후 이혼율 감소 가설에 대한 결과를 분석합니다.

model2의 결과입니다.



```
Call:
lm(formula = 이혼율 ~ 연령별, data = df_age[df_age$연령별 >
    4044, ])

Residuals:
        7         8        10        11        12
 0.620448  1.444818 -2.554342 -0.612325  0.006162
       13
 1.095238

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.706e+02  4.521e+00   37.73 2.95e-06 ***
연령별       -2.161e-02  7.695e-04  -28.08 9.57e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.626 on 4 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.9937
F-statistic: 788.5 on 1 and 4 DF,  p-value: 9.569e-06
```

결과를 해석해보겠습니다.

회귀계수 (Coefficients)는 (Intercept): 회귀식의 절편은 170.6이며, 이 값은 유의미한 값을 가집니다 (p-value < 0.001). 이는 연령이 0일 때 이혼율의 추정값이 170.6임을 의미합니다.
연령별 회귀식의 기울기는 -0.02161이며, 이 값은 유의미한 값을 가집니다 (p-value < 0.001). 이는 연령이 증가함에 따라 이혼율이 감소하는 경향을 나타냅니다.
유의성 검정 (Significance)입니다.

모든 회귀계수의 p-value가 매우 작으며, '***' 기호로 표시됩니다. 이는 모든 회귀계수가 통계적으로 유의미하다는 것을 나타냅니다 (p-value < 0.001).

적합도 분석 (Goodness of Fit)입니다.

Multiple R-squared (다중 결정 계수)는 0.995로 매우 높은 값을 가집니다. 이는 회귀식이 데이터를 매우 잘 설명한다는 것을 의미합니다.
Adjusted R-squared (조정된 다중 결정 계수)는 0.9937로, 변수의 수와 데이터의 크기를

고려한 적합도 지표입니다.
F-statistic (F-통계량)은 788.5이며, 이 값은 매우 큰 값을 가집니다. 이는 회귀식이 통계적으로 유의미하다는 것을 나타냅니다 (p-value < 0.001).

따라서, model2의 결과를 종합적으로 해석하면 다음과 같습니다.

45세 이후부터 연령이 증가함에 따라 이혼율은 감소하는 경향을 보입니다. 모델은 데이터를 매우 잘 설명하고, 회귀계수들은 통계적으로 유의미합니다.
모델의 적합도 지표인 다중 결정 계수와 F-통계량은 매우 높은 값을 가지므로, 회귀식이 통계적으로 유의미하다고 할 수 있습니다.

다음으로 model2의 모델링과 예측 가능성을 판별해 보겠습니다.

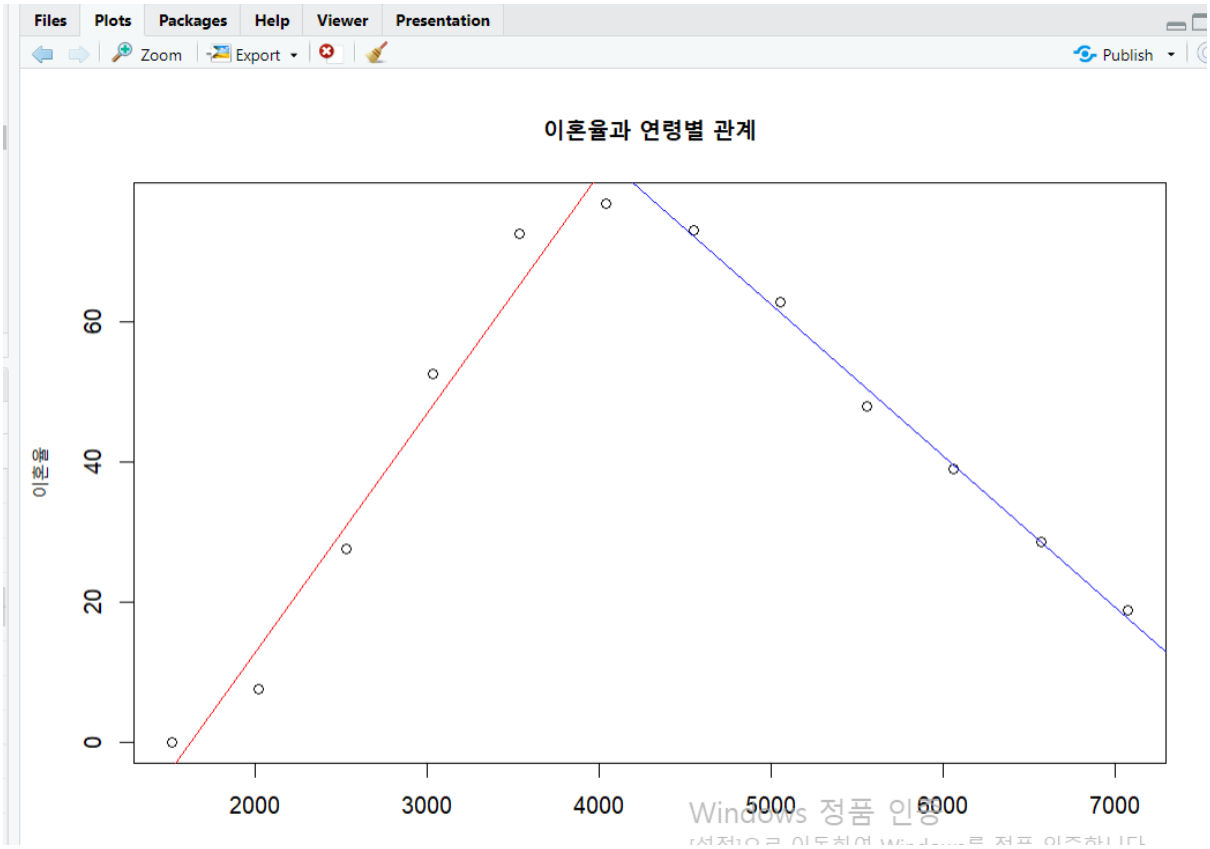
앞서 살표본 바와 같이 model2은 다음과 같은 선형 회귀 모델입니다.

이혼율 = 170.6 - 0.02161 * 연령별

이 모델은 연령별과 이혼율 간의 선형 관계를 나타내는 회귀식입니다.

아래는 가능성을 판별하기 위한 코드입니다.

```
# 데이터 탐색
plot(df_age$연령별, df_age$이혼율, main = "이혼율과 연령별 관계", xlab = "연령별", ylab = "이혼율")
# 모델2 시각화
abline(model2, col = "blue")
```



시각화 결과 이혼률과 model1(45세 이전까지의 이혼률), model2(45세 이후 이혼률) 간에 선형적인 관계가 있음을 확인할 수 있습니다.

따라서 우리는 데이터에서 model1과 model2를 통해 예측하는 것이 가능하며, 선형 회귀 모델이 적절한 모델링 방법일 수 있습니다.

데이터 취득, 정제 ,가공 시각화

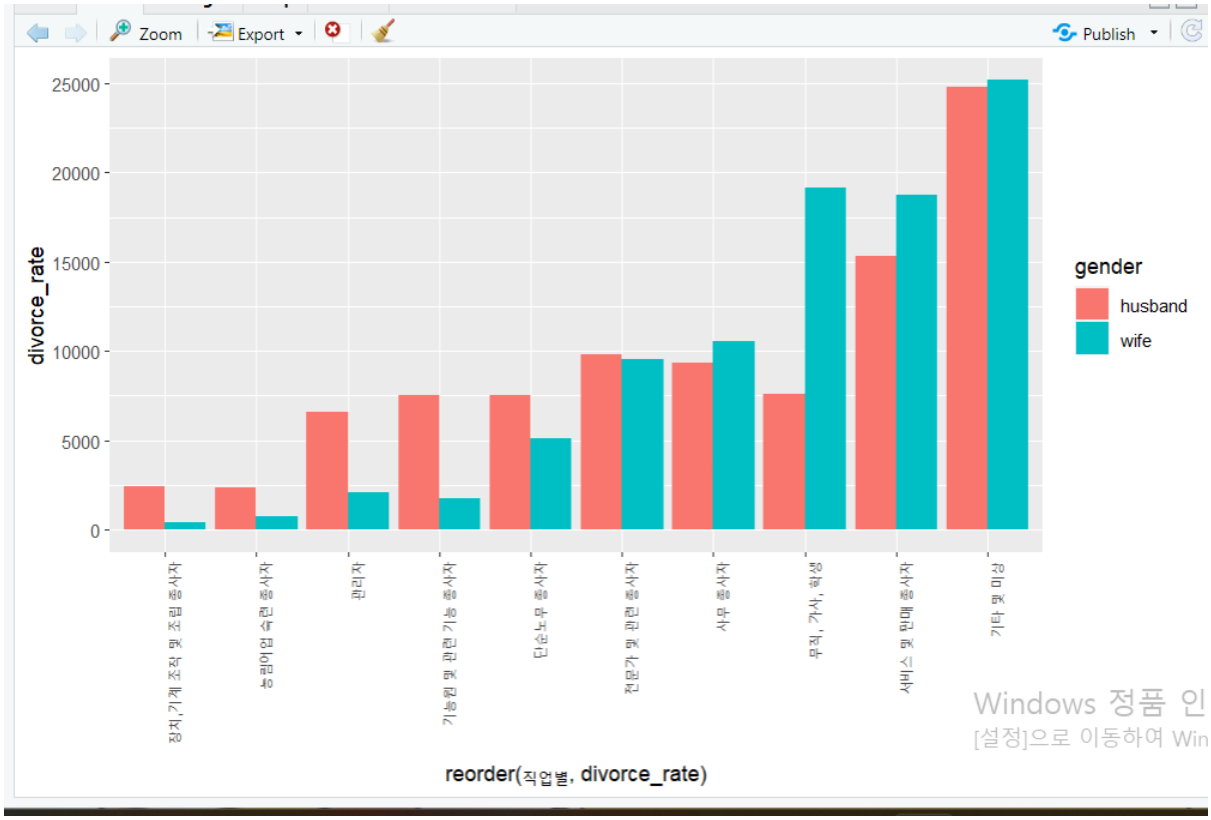
이혼을 결정하는 변수가 무엇일까 찾아보았습니다. 또 다른 유의미한 변수는 **직업**이었습니다..

출저: 통계청,「인구동향조사」, 2022, 2023.05.08, 시도/직업별(2008~) 이혼
https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B85029&conn_path=I2
.

시도별	직업별	2022	
		남편	아내
전국	계	93,232	93,232
	관리자	6,611	2,103
	전문가 및 관련 종사자	9,808	9,549
	사무 종사자	9,346	10,540
	서비스 및 판매 종사자	15,302	18,746
	농림어업 숙련 종사자	2,353	709
	기능원 및 관련 기능 종사자	7,507	1,763
	장치,기계 조작 및 조립 종사자	2,383	379
	단순노무 종사자	7,525	5,075
	무직, 가사, 학생	7,588	19,185
	기타 및 미상	24,809	25,183
서울특별시	계	13,174	13,516
	관리자	864	320

시도별	직업별	2022	
		남편	아내
	전문가 및 관련 종사자	1,407	1,387
	사무 종사자	1,573	1,557
	서비스 및 판매 종사자	1,926	2,126
	농림어업 숙련 종사자	22	21
	기능원 및 관련 기능 종사자	588	165
	장치,기계 조작 및 조립 종사자	94	12
	단순노무 종사자	644	401
	무직, 가사, 학생	1,047	2,343
	기타 및 미상	5,009	5,184
부산광역시	계	5,523	5,690

외 전국 모든 시 및 도 정보 포함



아래는 데이터를 **시각화** 하기 위해 **취득과 정제 가공의 과정 거친 코드입니다.**

```
# 필요한 패키지 로드
library(ggplot2)
library(dplyr)
library(tidyr)

# 데이터 불러오기
tb<-read_csv("시도_직업별_2008__이혼
_20230508094338.csv",locale=locale("ko",encoding="EUC-KR"),na=".")
df<-tb[-2,]  # 행 제거

# 행 이름 변경
df <- rename(df, "husband" = `2022...3`, "wife" = `2022...4`)

# 시도별 제거
df <- df %>% filter(시도별 != "시도별")
```

```
# 숫자형으로 변환
df$husband <- as.numeric(df$husband)
df$wife <- as.numeric(df$wife)

# 데이터 가공
df <- pivot_longer(df, cols=c("husband", "wife"), names_to = "gender", values_to =
"divorce_rate")
df <- df[order(df$직업별), ]  # 직업별 오름차순으로 정렬
df <- df %>% filter(직업별 != "계")



# 시각화
ggplot(df, aes(x=reorder(직업별, divorce_rate), y=divorce_rate, fill=gender)) +
  geom_col(position="dodge") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
  ---------------------------------------------------------------------------------------
```

먼저, 데이터 정제 과정에서 수행한 작업에 대해 설명드리겠습니다.

행 제거:
데이터 프레임 df에서 tb[-2,]를 사용하여 두 번째 행을 제거하였습니다. 이는 데이터 파일에서 변수명이 아닌 값을 가지고 있는 행을 제거하는 작업입니다.

행 이름 변경:
rename 함수를 사용하여 변수 이름을 "2022...3"을 의미하는 "husband"로, "2022...4"를 의미하는 "wife"로 변경하였습니다. 이는 변수명을 이해하기 쉽고 명확하게 변경하기 위한 작업입니다.

시도별 제거:
filter 함수를 사용하여 "시도별"이라는 값을 가지고 있는 행을 제거하였습니다. "시도별"은 직업별 데이터를 분석하는 데에는 필요하지 않은 값이므로 제거하였습니다.

숫자형으로 변환:
as.numeric 함수를 사용하여 "husband"와 "wife" 변수를 숫자형으로 변환하였습니다. 이

는 현재 문자열로 표현된 값들을 숫자로 변환하여 분석에 활용하기 위한 작업입니다.

다음으로, 데이터 가공 과정에 대해 설명드리겠습니다.

pivot_longer 함수를 사용하여 데이터 형태 변환:
pivot_longer 함수를 사용하여 "husband"와 "wife" 변수를 "gender"와 "divorce_rate" 변수로 변환하였습니다. 이는 데이터를 '긴 형식'으로 변환하여 직업별 이혼율을 비교하기 쉽게 만들기 위한 작업입니다. "gender" 변수는 성별을, "divorce_rate" 변수는 해당 성별의 이혼율을 나타냅니다.

직업별 오름차순으로 정렬:
df를 "직업별" 변수를 기준으로 오름차순으로 정렬하였습니다. 이는 직업별로 데이터를 시각화할 때 순서를 보장하기 위한 작업입니다.

전체 시도를 계산하는 계열을 지워 시각회 시 의미없는 열을 삭제할 수 있도록 하였습니다.

마지막으로, 시각화 과정에 대해 설명드리겠습니다.

ggplot 함수를 사용하여 그래프 생성:
ggplot 함수를 사용하여 그래프를 생성하였습니다. 데이터프레임 df를 기반으로 그래프를 구성할 것을 명시합니다.

aes 함수를 사용하여 변수 지정:
aes 함수를 사용하여 x축에는 "직업별" 변수를 재정렬하여 표시하고, y축에는 "divorce_rate" 변수를 사용하도록 설정합니다. fill 옵션을 통해 "gender" 변수를 색상으로 구분합니다.

geom_col 함수를 사용하여 막대 그래프 생성:
geom_col 함수를 사용하여 막대 그래프를 생성합니다. position = "dodge" 옵션을 통해 성별에 따라 막대를 나란히 표시합니다.

theme 함수를 사용하여 x축 레이블 설정:
theme 함수와 axis.text.x 옵션을 사용하여 x축 레이블을 90도 회전하여 표시하고, 텍스트의 위치를 조정합니다. 이는 x축에 많은 라벨이 있을 경우 라벨이 겹치지 않도록 하기

위한 작업입니다.

이렇게 데이터 정제, 데이터 가공, 데이터 시각화 과정을 거친 이유는 직업별 이혼율을 성별로 구분하여 비교하고 시각적으로 표현하기 위함입니다. 데이터를 정제하고 가공함으로써 분석 목적에 부합하는 형태로 데이터를 준비하였으며, 시각화를 통해 직업과 성별에 따른 이혼율의 차이를 한눈에 알아볼 수 있도록 하였습니다.

그래프에서 알 수 있듯이 기타 및 미상 즉 직업이 없거나 서비스 및 판매 종사자와 같이 직업의 안정성이 떨어지는 직업에선 이혼율이 높았고 관리자 및 장치 기계 조작 조립 종사자와 같이 안정성이 다소 높은 직업에서 이혼률이 낮아지는 것을 볼 수 있습니다.

가설검정과 선형회귀, 모델링과 예측 가능성

가설 검정을 시행하기 위해서 일단 각 직업별 연봉 정보를 조사하였습니다.

관리자: 약 5,000만 원 - 1억 원
전문가 및 관련 종사자: 약 3,000만 원 - 8,000만 원
사무 종사자: 약 2,000만 원 - 4,000만 원
서비스 및 판매 종사자: 약 1,500만 원 - 4,000만 원
농림어업 숙련 종사자: 약 2,000만 원 - 4,000만 원
기능원 및 관련 기능 종사자: 약 2,000만 원 - 5,000만 원
장치, 기계 조작 및 조립 종사자: 약 1,800만 원 - 3,500만 원
단순노무 종사자: 약 1,500만 원 - 2,500만 원
무직, 가사, 학생: 수입이 없거나 미상
기타 및 미상: 다양한 경우가 있어 평균 연봉 범위를 정확히 파악하기 어렵습니다

이는 통계청, 인력 관리 기관, 산업 단체, 인터넷 취업 사이트, 급여 조사 기관 등의 신뢰할 수 있는 정보원을 참고하였습니다.

위의 조사한 연봉 정보를 바탕으로 열을 생성하기 위해 다음과 같이 연봉을 설정합니다.

직업군 평균연봉
1 장치,기계 조작 및 조립 종사자 25000000
2 농림어업 숙련 종사자 20000000
3 관리자 60000000
4 기능원 및 관련 기능 종사자 35000000
5 단순노무 종사자 15000000
6 사무 종사자 40000000
7 서비스 및 판매 종사자 30000000
8 전문가 및 관련 종사자 50000000
9 무직, 가사, 학생 10000000
10 기타 및 미상 12000000

아래는 코드입니다.

```
# 필요한 패키지 로드
library(ggplot2)
library(dplyr)
library(tidyr)

# 데이터 불러오기
tb<-read_csv("시도_직업별_2008__이혼
_20230508094338.csv",locale=locale("ko",encoding="EUC-KR"),na=".")
df<-tb[-2,]  # 행 제거

# 행 이름 변경
df <- rename(df, "husband" = `2022...3`, "wife" = `2022...4`)

# 시도별 제거
df <- df %>% filter(시도별 != "시도별")

# 숫자형으로 변환
df$husband <- as.numeric(df$husband)
df$wife <- as.numeric(df$wife)
```

```
# 데이터 가공
df <- pivot_longer(df, cols=c("husband", "wife"), names_to = "gender", values_to =
"divorce_rate")
df <- df[order(df$직업별), ]  # 직업별 오름차순으로 정렬
df <- df %>% filter(직업별 != "계")



# 시각화
ggplot(df, aes(x=reorder(직업별, divorce_rate), y=divorce_rate, fill=gender)) +
  geom_col(position="dodge") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

-------------------------------------------------------------------------------
# 직업별 평균 연봉 데이터 추가
avg_salary <- c(60000000, 50000000, 40000000, 30000000, 20000000, 35000000,
25000000, 15000000, 10000000, 12000000)

# 데이터 프레임의 행 개수와 평균 연봉 데이터 개수 일치시키기
avg_salary <- rep(avg_salary, each = nrow(df) / length(avg_salary))

# 평균 연봉 열 추가
df$avg_salary <- avg_salary

# 가설검정
# 직업군별 평균 연봉과 이혼률 간에 상관관계가 있는지 검정
correlation <- cor.test(df$avg_salary, df$divorce_rate)
print(correlation)

# 선형회귀
# 직업군별 평균 연봉을 기준으로 이혼률을 예측하는 선형회귀 모델 구축
model <- lm(divorce_rate ~ avg_salary, data = df)
summary(model)

# 모델 시각화
# 회귀선 그리기
```

```
plot(df$avg_salary, df$divorce_rate, xlab = "평균 연봉", ylab = "이혼률")
abline(model, col = "red")

# 데이터 탐색
# 산점도 그리기
plot(df$avg_salary, df$divorce_rate, xlab = "평균 연봉", ylab = "이혼률")



# 모델링과 예측 가능성 판단
# 예측 모델의 성능을 평가하고 예측 가능성을 판단하는 작업 수행
predictions <- predict(model, newdata = df)
```

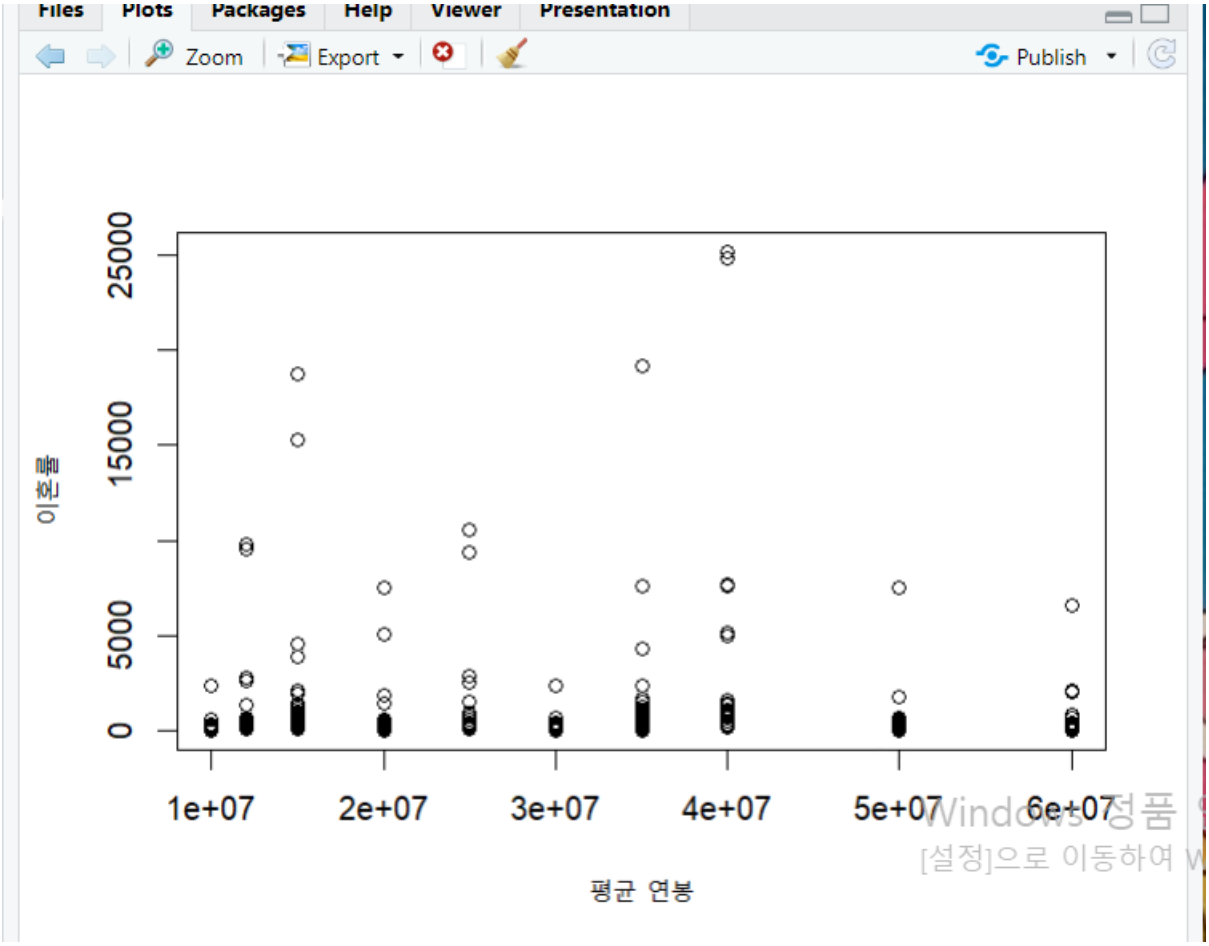
위의 코드는 avg_salary라는 변수에 직업별 평균 연봉 데이터를 저장합니다. avg_salary 변수에는 각 직업군에 대한 평균 연봉을 순서대로 나열한 벡터가 저장되어 있습니다. 이후 rep 함수를 사용하여 데이터 프레임의 행 개수와 평균 연봉 데이터 개수를 일치시킵니다. 이렇게 하면 데이터 프레임의 행 개수와 평균 연봉 데이터 개수가 맞아지게 됩니다. 마지막으로 df$avg_salary에 avg_salary 값을 추가하여 직업별 평균 연봉 열을 데이터 프레임에 추가합니다.

또 cor.test 함수를 사용하여 직업군별 평균 연봉과 이혼률 간의 상관관계를 검정합니다. cor.test 함수는 두 변수 간의 상관계수를 계산하고, 이를 검정하여 상관관계의 유의성을 판단합니다. 위의 코드에서는 df$avg_salary와 df$divorce_rate를 입력으로 사용하여 상관관계를 계산하고, correlation 변수에 결과를 저장합니다. 마지막으로 print(correlation)을 통해 상관관계 검정 결과를 출력합니다.

lm 함수를 사용하여 직업군별 평균 연봉을 기준으로 이혼률을 예측하는 선형회귀 모델을 구축합니다. lm 함수는 선형회귀 모델을 생성하기 위해 사용됩니다. 위의 코드에서는 divorce_rate를 종속 변수로, avg_salary를 독립 변수로 지정하여 모델을 구축하고, model 변수에 결과를 저장합니다. summary(model)을 통해 모델의 요약 정보를 출력합니다.

plot 함수와 abline 함수를 사용하여 모델 시각화를 수행합니다. plot 함수를 통해 산점도를 그리고, abline 함수를 사용하여 회귀선을 그립니다. xlab과 ylab 인자를 사용하여 x축과 y축의 레이블을 지정하고, col 인자를 사용하여 회귀선의 색상을 지정합니다.

데이터 탐색을 위해 산점도를 그립니다. plot 함수를 사용하여 df$avg_salary를 x축, df$divorce_rate를 y축으로 하는 산점도를 그립니다. xlab과 ylab 인자를 사용하여 x축과 y축의 레이블을 지정합니다.



시각화 자료.

```
> print(correlation)

        Pearson's product-moment correlation

data:  df$avg_salary and df$divorce_rate
t = -0.00773, df = 378, p-value = 0.9938
alternative hypothesis: true correlation is not equal
to 0
95 percent confidence interval:
 -0.1009954  0.1002082
sample estimates:
          cor
-0.000397589
```

상관계수

```
> summary(model)

Call:
lm(formula = divorce_rate ~ avg_salary, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
 -981.7  -883.9  -675.4  -298.3 24202.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.834e+02  3.000e+02   3.278  0.00114 **
avg_salary  -6.893e-08  8.918e-06  -0.008  0.99384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 2748 on 378 degrees of freedom
Multiple R-squared:  1.581e-07, Adjusted R-squared:  -0.00264
5
F-statistic: 5.975e-05 on 1 and 378 DF,  p-value: 0.9938

>
```

summary결과

가설검정 결과인 correlation의 출력과 선형회귀 모델의 결과인 summary(model)의 출력

을 해석해보겠습니다:

가설검정 결과 해석 (correlation)입니다.

Pearson의 상관계수 (Pearson's product-moment correlation)를 이용한 상관분석 결과입니다.

데이터셋 df의 avg_salary 변수와 divorce_rate 변수 사이의 상관계수를 계산한 결과입니다.
검정 통계량 (t-statistic)은 -0.00773이고, 자유도 (df)는 378입니다.
p-value (유의확률)는 0.9938로 나왔습니다.
대립가설 (alternative hypothesis)은 "상관계수가 0이 아니다"로 설정되었습니다.
95% 신뢰구간 (confidence interval)는 -0.1009954부터 0.1002082까지입니다.
표본 추정치 (sample estimate)인 상관계수 (cor)는 -0.000397589입니다.
상관계수는 -0.000397589로 매우 낮으며, p-value가 0.9938로 매우 높습니다. 이는 avg_salary와 divorce_rate 사이에 유의미한 상관관계가 없다는 결론을 내릴 수 있습니다. 따라서, 직업군별 평균 연봉과 이혼률 간에는 통계적으로 유의한 상관관계가 없다고 할 수 있습니다.

선형회귀 결과 해석 (summary(model))입니다.

선형회귀 모델의 결과를 나타냅니다.
종속변수인 divorce_rate를 독립변수인 avg_salary로 예측하는 모델입니다.
회귀식은 divorce_rate = 983.4 - 6.893e-08 * avg_salary입니다.
avg_salary의 계수 (coefficient)는 -6.893e-08이고, 해당 계수의 표준오차 (standard error)는 8.918e-06입니다.
t-value는 -0.008이고, p-value는 0.99384입니다.
회귀 모델의 설명력을 나타내는 결정계수 (Multiple R-squared)는 1.581e-07이고, 조정된 결정계수 (Adjusted R-squared)는 -0.002645입니다.
F-통계량 (F-statistic)는 5.975e-05이며, 이 모델의 p-value는 0.9938입니다.
해석: 선형회귀 모델의 결과를 종합해보면, avg_salary 변수는 divorce_rate 변수를 예측하는 데에 유의미한 영향을 미치지 않는 것으로 나타납니다. avg_salary의 계수가 매우 작고, p-value가 0.99384로 매우 높기 때문에, 직업군별 평균 연봉을 기준으로 이혼률을 예측하는데에는 유의미한 예측력을 가지지 못한다고 할 수 있습니다. 즉, 이 모델은 avg_salary와 divorce_rate 사이의 관계를 설명하기에는 적절하지 않은 것으로 나타납니

다.

따라서 귀무가설(H0): 직업별 연봉이 높으면 이혼율이 낮을 것이다.
대립가설(H1): 직업별 연봉이 높으면 이혼율이 낮지 않을 것이다.
유의수준(alpha): 예를 들어, 0.05로 설정합니다. (p-value가 유의수준보다 작으면 귀무가설 기각)이라고 설정했다고 해도

상관분석 결과와 선형회귀 모델 결과를 종합해보면, 직업별 연봉과 이혼율 사이에는 통계적으로 유의미한 관계가 없습니다.

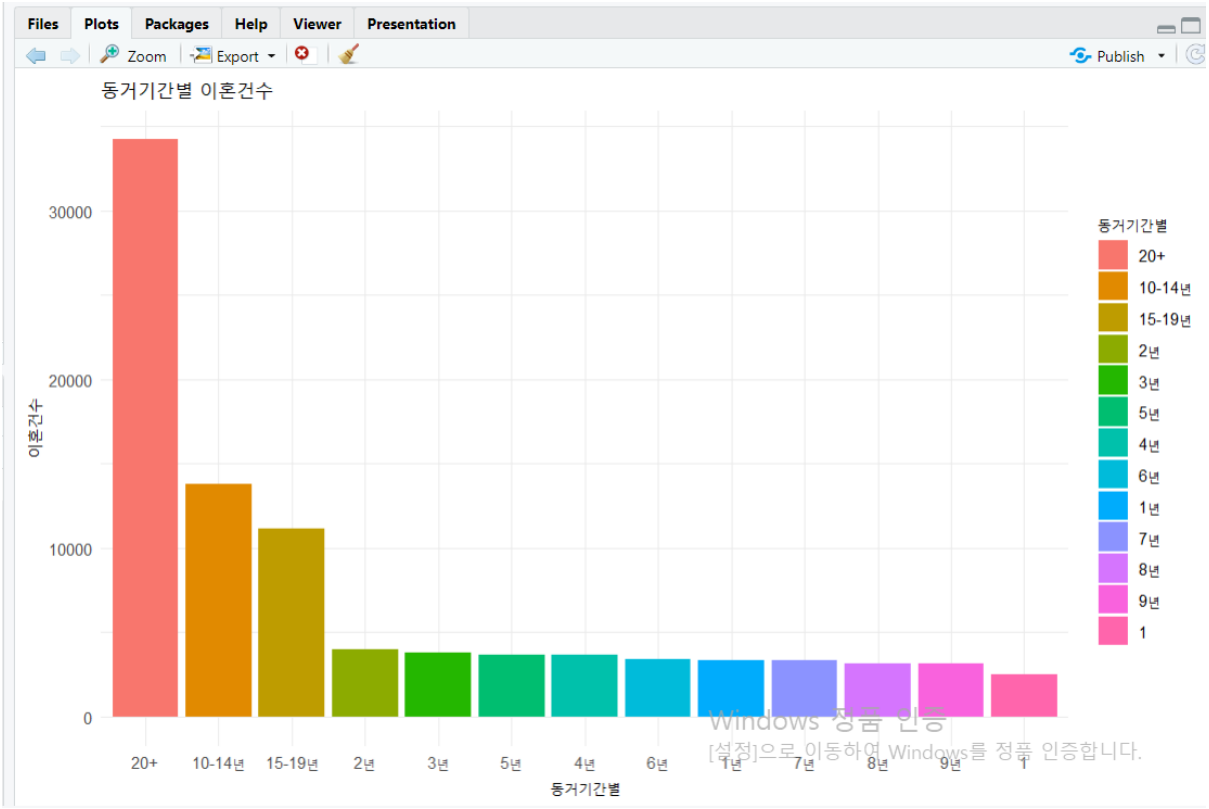
따라서, 직업별 연봉을 기준으로 이혼율을 예측하는 모델은 적절하지 않습니다.
.

데이터 취득, 정제 ,가공 시각화

1. 이혼이 유발 가능한 또 하나의 변수를 찾았습니다. 바로 동거 기간 입니다.

해당 데이터의 출저를 밝힙니다.
주소: https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B85008&conn_path=I2
출저: 통계청,「인구동향조사」, 2022, 2023.05.17, 연령(5세)/ 혼인지속기간(동거기간)별 이혼

아래 막대그래프는 동거 기간별 이혼 건수를 시각화 한 것 입니다.



데이터를 **시각화** 하기 위해 **취득과 정제 가공의 과정 거친 코드입니다.**

```
# 필요한 라이브러리 로드
library(tidyverse)
library(ggplot2)
library(dplyr)

# 데이터 취득 및 정제
data <- read_csv("연령_5세___혼인지속기간_동거기간_별_이혼
_20230517123720.csv",locale=locale("ko",encoding="EUC-KR"),na=".")
data
data <- data[-1, ]  # 첫 번째 행 제거 (헤더 제거)
data$동거기간별 <- gsub("년 미만", "", data$동거기간별)  # "년 미만" 문자열 삭제
data$동거기간별 <- gsub("~", "-", data$동거기간별)  # "~" 문자열을 "-"로 변경
data$동거기간별 <- gsub("년 이상", "+", data$동거기간별)  # "년 이상" 문자열을 "+"로
변경
data$동거기간별 <- as.factor(data$동거기간별)  # 동거기간별을 팩터로 변환
```

```
# 데이터 가공
data <- data %>%
  mutate(이혼건수 = as.integer(이혼건수))  # 이혼건수를 정수로 변환

# 데이터 시각화
data <- data %>%
  mutate(동거기간별 = reorder(동거기간별, -이혼건수))  # 동거기간별을 이혼건수 기준으로 내림차순 정렬

ggplot(data, aes(x = 동거기간별, y = 이혼건수, fill = 동거기간별)) +
  geom_bar(stat = "identity") +
  labs(x = "동거기간별", y = "이혼건수", fill = "동거기간별") +
  ggtitle("동거기간별 이혼건수") +
  theme_minimal()
```

먼저 데이터 취득 및 정제의 과정입니다.

먼저, read_csv 함수를 사용하여 CSV 파일을 읽어옵니다. 파일의 인코딩은 EUC-KR로 설정하고, 결측치는 "."으로 처리합니다. 데이터를 읽어온 후, 첫 번째 행을 제거하여 헤더를 제거합니다. 이렇게 하는 이유는 분석에 필요한 데이터를 남기고 불필요한 헤더를 제거하여 데이터를 정제하는 것입니다.

다음으로 데이터 가공의 과정입니다.

가공 단계에서는 문자열을 처리하고, 변수의 형식을 변경합니다. gsub 함수를 사용하여 "년 미만", "~", "년 이상"과 같은 문자열을 변경합니다. as.factor 함수를 사용하여 동거기간별 변수를 팩터로 변환합니다. 이러한 데이터 가공 과정은 데이터를 더 적합한 형식으로 변환하고, 시각화 단계에서 데이터를 잘 표현하기 위한 작업입니다.

마지막으로 데이터 시각화의 과정입니다.

시각화 단계에서는 ggplot 함수를 사용하여 동거기간별 이혼건수를 시각화합니다. aes 함수를 사용하여 x축을 동거기간별, y축을 이혼건수로 설정하고, fill을 동거기간별로 설정합니다. geom_bar(stat = "identity")를 사용하여 막대 그래프를 그리고, labs 함수를 사용하여 x축, y축, 그리고 범례의 레이블을 설정합니다. 또한, ggtitle 함수를 사용하여 그래프의 제목을 설정합니다. theme_minimal 함수를 사용하여 그래프의 테마를 설정합니다.

이렇게 데이터를 가공하고 시각화함으로써 동거기간별 이혼건수를 파악할 수 있습니다. 시각화 결과를 통해 동거기간이 길어질수록 이혼건수가 증가하는 경향을 확인할 수 있습니다.

시각화 된 그래프에서 알 수 있듯이 동거 기간(결혼 연수 , 출저의 자료를 살펴보면 동거기간과 결혼 연수는 같은 의미로 사용됩니다.)이 늘어날수록 이혼 건수가 늘어나는 것을 확인할 수 있습니다.

가설검정과 선형회귀, 모델링과 예측 가능성

아래는 가설검정, 모델링을 수행하는 코드입니다.

```
# 필요한 라이브러리 로드
library(tidyverse)
library(ggplot2)
library(dplyr)

# 데이터 취득 및 정제
data <- read_csv("연령_5세___혼인지속기간_동거기간_별_이혼
_20230517123720.csv",locale=locale("ko",encoding="EUC-KR"),na=".")
data
data <- data[-1, ]  # 첫 번째 행 제거 (헤더 제거)
data$동거기간별 <- gsub("년 미만", "", data$동거기간별)  # "년 미만" 문자열 삭제
data$동거기간별 <- gsub("~", "-", data$동거기간별)  # "~" 문자열을 "-"로 변경
data$동거기간별 <- gsub("년 이상", "+", data$동거기간별)  # "년 이상" 문자열을 "+"로
변경
data$동거기간별 <- as.character(data$동거기간별)  # 동거기간별을 문자열로 변환
```

```
# 데이터 가공
data <- data %>%
  mutate(이혼건수 = as.integer(이혼건수))  # 이혼건수를 정수로 변환

# 데이터 시각화
data <- data %>%
  mutate(동거기간별 = reorder(동거기간별, -이혼건수))  # 동거기간별을 이혼건수 기준으로 내림차순 정렬

ggplot(data, aes(x = 동거기간별, y = 이혼건수, fill = 동거기간별)) +
  geom_bar(stat = "identity") +
  labs(x = "동거기간별", y = "이혼건수", fill = "동거기간별") +
  ggtitle("동거기간별 이혼건수") +
  theme_minimal()



# 동거기간별을 정수로 변환
data$동거기간별 <- gsub("₩₩+","",data$동거기간별)
data$동거기간별 <- gsub("₩₩-","",data$동거기간별)
data$동거기간별 <- gsub("년","",data$동거기간별)
data$동거기간별 <- gsub("1014","10",data$동거기간별)
data$동거기간별 <- gsub("1519","15",data$동거기간별)
data$동거기간별 <- as.integer(data$동거기간별)


# 선형 회귀 모델링
model <- lm(이혼건수 ~ 동거기간별, data = data)

# 모델 요약
summary(model)

#시각화
abline()
```

```
# 예측 가능성 평가
prediction <- predict(model, newdata = data)
accuracy <- cor(data$이혼건수, prediction)
accuracy

# 시각화 및 회귀 직선 그리기
ggplot(data, aes(x = 동거기간별, y = 이혼건수)) +
  geom_point() +
  geom_abline(intercept = coef(model)[1], slope = coef(model)[2], color = "red") +
  labs(x = "동거기간별", y = "이혼건수") +
  ggtitle("동거기간별 이혼건수") +
  theme_minimal()
```

동거기간별 변수를 정수로 변환합니다.

gsub 함수를 사용하여 "+"와 "-" 기호, "년" 문자를 삭제합니다.
"1014"를 "10"으로, "1519"를 "15"으로 대체합니다.
as.integer 함수를 사용하여 동거기간별 변수를 정수로 변환합니다.

선형 회귀 모델링합니다.

lm 함수를 사용하여 선형 회귀 모델을 생성합니다. 이혼건수를 종속 변수로, 동거기간별을 독립 변수로 지정합니다.
생성된 모델은 model 변수에 저장됩니다.

모델 요약합니다.

summary 함수를 사용하여 모델의 요약 통계 정보를 출력합니다.

예측 가능성을 평가합니다.

생성된 모델을 사용하여 동일한 데이터에 대한 예측값을 계산합니다.
실제 이혼건수와 예측값 간의 상관관계를 계산하여 예측 가능성을 측정합니다. 계수 값을 accuracy 변수에 저장합니다.

시각화 및 회귀 직선을 그립니다.

ggplot 함수를 사용하여 데이터를 시각화합니다. x 축은 동거기간별, y 축은 이혼건수로 설정합니다.
geom_point 함수를 사용하여 점을 그립니다.
geom_abline 함수를 사용하여 회귀 직선을 그립니다. 회귀 모델의 절편과 기울기를 사용하며, 색상은 빨강으로 지정합니다.
축 레이블과 제목을 설정하고, 시각화의 테마를 minimal로 지정합니다.

가설 검정 단계입니다.

귀무가설(H0): 동거기간별과 이혼건수 간에는 유의미한 관계가 없다.
대립가설(H1): 동거기간별 변수는 이혼건수에 영향을 미친다.

유의수준은 0.01(1%)을 사용합니다.

```
>
> # 모델 요약
> summary(model)

Call:
lm(formula = 이혼건수 ~ 동거기간별, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6673.6 -3842.7   449.1  3241.9  9919.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2046.9     2322.7  -0.881 0.397035
동거기간별    1316.9      263.4   5.000 0.000403 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5094 on 11 degrees of freedom
Multiple R-squared:  0.6944,    Adjusted R-squared:  0.6667
F-statistic:    25 on 1 and 11 DF,  p-value: 0.0004025

>
```

선형 회귀 모델은 이혼건수를 동거기간별로 설명하고 있습니다. 분석 결과에 따르면, 모델의 설명력을 나타내는 R-squared 값은 0.6944이며, 조정된 R-squared 값은 0.6667입니다. 이는 동거기간별이 이혼건수의 변동을 약 69.44% 설명하고 있다는 의미입니다.

동거기간별 변수의 추정 계수는 1316.9이고, 이는 통계적으로 유의미합니다. 동거기간별 변수의 p-value는 0.000403로, 유의수준 0.05에서 유의미한 관련성을 나타냅니다. 따라서 동거기간별 변수는 이혼건수에 유의한 영향을 미친다고 할 수 있습니다.

F-statistic는 25이며, 이는 동거기간별 변수의 유의성을 평가하는데 사용되는 검정 통계량입니다. p-value는 0.0004025로, 유의수준 0.05에서 동거기간별 변수가 모델에 유의한 기여를 한다는 것을 나타냅니다.

모델의 잔차(오차)에 대한 정보도 제공됩니다. 잔차의 최소값은 -6673.6이고, 최대값은 9919.1입니다. 이는 모델이 실제 데이터와 얼마나 잘 일치하는지를 나타내는 지표입니다. 따라서, 이 분석 결과를 바탕으로 동거기간별 변수는 이혼건수에 유의한 영향을 미치며, 모델은 이혼건수의 변동을 약 69.44% 설명할 수 있다고 할 수 있습니다.

분석 결과를 토대로 귀무가설을 평가해보겠습니다.

귀무가설(H0): 동거기간별 변수는 이혼건수에 영향을 미치지 않는다.

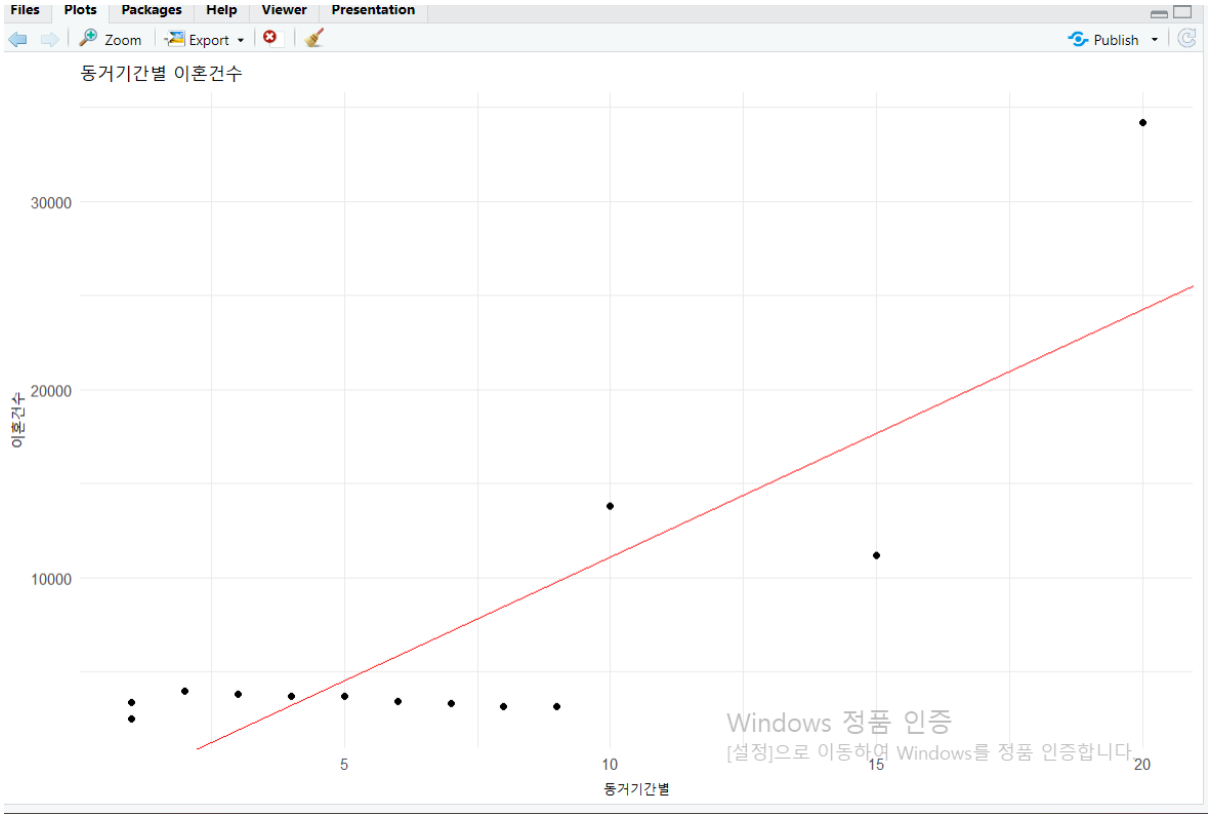
대립가설(H1): 동거기간별 변수는 이혼건수에 영향을 미친다.

분석 결과를 통해 동거기간별 변수의 추정 계수는 유의미하게 나타났고, p-value는 0.000403로 매우 낮게 나타났습니다. 또한, 모델의 R-squared 값도 0.6944로 이혼건수의 변동을 상당 부분 설명하고 있음을 나타냈습니다. 이러한 결과를 고려할 때, 귀무가설을 기각하고 대립가설을 채택하는 것이 적절할 것입니다. 즉, 동거기간별 변수는 이혼건수에 영향을 미친다고 할 수 있습니다.

모델링과 예측 가능성 판별:

코드에서는 선형 회귀 모델을 사용하여 동거기간별과 이혼건수 간의 관계를 모델링하고 있습니다. 이를 통해 동거기간별이 이혼건수에 어떤 영향을 미치는지 파악할 수 있습니다. 모델 요약 결과를 확인하여 회귀 계수와 통계적 유의성을 평가할 수 있습니다.
예측 가능성은 모델을 사용하여 새로운 데이터에 대한 이혼건수를 예측할 수 있는지를

평가합니다. 코드에서는 모델을 사용하여 예측을 수행하고, 예측값과 실제값 간의 상관관계를 계산하여 예측의 정확성을 측정합니다. 이를 통해 모델의 예측 가능성을 평가할 수 있습니다.



선형회귀 시각화 자료

```
> # 예측 가능성 평가
> prediction <- predict(model, newdata = data)
> accuracy <- cor(data$이혼건수, prediction)
> accuracy
[1] 0.833334
>
> # 시각화 및 회귀 직선 그리기
> ggplot(data, aes(x = 동거기간별, y = 이혼건수)) +
+   geom_point() +
+   geom_abline(intercept = coef(model)[1], slope = coef(model)[2], color = "red") +
+   labs(x = "동거기간별", y = "이혼건수") +
+   ggtitle("동거기간별 이혼건수") +
+   theme_minimal()
>
```

예측한 결과와 실제 이혼건수 간의 상관 관계를 확인하기 위해 예측값(prediction)과 실제값(data$이혼건수) 사이의 상관 계수를 계산하였습니다. 계산 결과, 상관 계수는 0.833334로 나타났습니다.

상관 계수는 -1부터 1까지의 범위를 가지며, 0에 가까울수록 예측과 실제 값 사이에 상관 관계가 없음을 나타내고, 1에 가까울수록 강한 양의 상관 관계가 있음을 나타냅니다.

따라서, 상관 계수 0.833334는 예측값과 실제값 사이에 강한 양의 상관 관계가 있다는 것을 의미합니다. 이는 모델이 이혼건수를 상당히 정확하게 예측할 수 있음을 나타내며, 모델의 예측 가능성이 높다는 것을 의미합니다.

결론

본 프로젝트는 이혼 여부를 종속 변수로 사용하여 데이터 분석을 수행하였습니다. 데이터를 획득하고 정제한 후, 변수들 간의 상관 계수를 계산하여 상관 관계를 파악하고, 이를 시각화하여 분석한 결과를 도출하였습니다. 또한, 가설 검정을 통해 변수들 간의 관계를 통계적으로 검증하고, 선형회귀 모델을 활용하여 이혼 여부를 예측하는 모델을 학습시켰습니다.

분석 결과, 시도별, 연령별, 혼인 기간별로는 선형회귀가 잘 이루어져 이혼 여부와 다양한 요인들 간의 관계를 설명할 수 있었습니다. 이를 통해 이혼율과 관련된 패턴이나 추세를 파악할 수 있었습니다.

하지만 직업별로는 설득력있는 유의미한 결과를 얻지 못했습니다. 이는 다른 요인들이 직업별 이혼률에 큰 영향을 미치지 않는다는 것을 의미할 수 있습니다. 따라서, 직업별로 이혼률을 예측하는 모델을 개발하는 데는 한계가 있을 수 있습니다.

본 프로젝트는 이혼에 영향을 미치는 요인들을 이해하고, 이혼 여부를 예측하는 데 도움이 되는 통계적인 분석과 모델링을 수행함으로써, 사회 현상을 이해하고 개인 및 정책 수립에 기여할 수 있는 정보를 제공하고자 합니다.

결론적으로, 이 프로젝트를 통해 이혼율과 관련된 다양한 요인들 사이의 상관 관계를 파악하고, 이를 통해 이혼 여부를 예측하는 모델을 개발하였습니다. 이를 통해 개인들은 이혼에 대한 위험 요인들을 파악하고 대비할 수 있으며, 정부 및 사회 기관은 이혼율을 낮추기 위한 정책 수립에 참고 자료로 활용할 수 있을 것입니다.