

④ 경테분 2 스텝 고도정리

- ① 데이터파일이름 ← read_csv("file.csv") : file을 테이터파일 이름으로 저장하고 불러오기
- ② 우선 library(readr) 먼저하기
- ③ str(데이터 아 데이터 \$변수) : 데이터 또는 데이터의 변수에 대해 다양한 정보 (선택, 개수, 예시...)
- ④ glimpse(데이터) : 데이터의 다양한 정보를 보여줌 - str보다 조금 깔끔하게
- ⑤ 우선 library(dplyr) 먼저하기
- ⑥ summary(데이터) : 기초통계량들을 보여줌, 즉 문자형식도 제외하고 계량형식도 (int, num...)의 수가 적인 경우
- ⑦ 결속치의 개수 확인 (NA)
- ⑧ freq(데이터 \$변수) : 해당변수의 빈도수와 악대그래프
- ⑨ 우선 library(describe) 먼저하기 - 빙수변수 가능
- ⑩ 데이터 %>% group_by(변수) %>% summarise(별명=mean or var or sd... (변수))
 - ! 변수별로 그룹을 나누고 기본통계량을 조건에 맞게 제작
- ⑪ 우선 library(dplyr) 먼저하기 ⑫ %>% 항수들 핀다
- ⑬ mean / var / sd(데이터 \$변수, na.rm=T) : 데이터 변수에서 결속치(NA)를 제외하고 계산

- ⑭ boxplot(데이터 \$변수) : 해당변수의 상자그라프, 단 변수는 계량형식도 (이상치 check 가능)
- ⑮ 데이터 \$변수 ← as.factor(데이터 \$변수) : 해당변수 칙도를 범주형으로 변형
- ⑯ 데이터 \$변수 ← as.POSIXct(데이터 \$변수) : 해당변수 칙도를 시간형으로 변형
- ⑰ table(데이터 \$변수) : 해당변수의 빈도수
- ⑱ 데이터 \$변수 ← factor(데이터 \$변수, levels=c("a", "b", "c", ...))
 - ! a, b, c 순으로 변수를 배열하고 칙도를 범주형으로 바꾼뒤 해당데이터의 변수에 지정
- ⑲ 데이터 \$변수 ← Tfeise(데이터 \$변수 = "a", "b", 데이터 \$변수) : 만약 데이터의 변수값이 a면 b로 바꾸고 아니면 데이터의 원래 변수값 그대로 두어라
- ⑳ T\$NA(데이터 \$변수) : 데이터의 변수가 결속치(NA)이다.
- ⑳ 데이터 ← 데이터 %>% drop_na() : 결속치(NA)가 있는 행들을 삭제
- ⑷ 우선 library(psych) 먼저하기

- ① descTP+TVE ← describe(데이터) : 해당데이터들을 계량적으로 계산한 descTP+TVE에 할당 (mean, sd, min, max...)
- ② 우선 library(psych) 먼저하기
- ③ descTP+TVE ← descTP+TVE %>% mutate(CL = mean-3*sd, UL = mean+3*sd)
- ④ 데이터 ← 데이터 %>% filter(변수 < K) (UL값)
 - ! 위에서 생성한 UL 값보다 작은 값을 안 filter 이용해서 추출 (UL보다 큰 값은 이상치)
- ⑤ 데이터 ← 데이터 %>% mutate(CL = mean-3*sd, UL = mean+3*sd)

회귀분석과 정

→ 완전하게 짜자에서 (불속자기화)

- 0 corr.test(데이터 [k, n], method="Pearson", alpha=0.05, use="pairwise.complete.obs")
(k가 7열이면 데이터의 전체 행을 의미하고 행에서 CC(7:14) → 7열~14열까지만으로 표시 가능)
우선 library(psych) 먼저하기
상관분석을 실행하는데 데이터에서 k열 n행끼리, 방법은 피어슨으로 실행
→ '+'로 여러 변수나 행 가능

- 0 테이터 저장법 ← lm(DV ~ IV) data=데이터 : 다중회귀식 수립하기
- 0 plot() : 도출한 4개의 그레프로 선형성, 정규성, 동반성 확인
- 0 ks.test(데이터 \$변수, pnorm, mean=mean(데이터 \$DV), sd=sd(데이터 \$DV)) :
데이터의 종속변수에서 정규분포를 그리는지 검정, 평균과 표준편차 이용
- 0 shapiro.test(데이터 \$변수) : N이 적을 때의 데이터의 종속변수의 정규성 검정도
- 0 htest(데이터 \$변수, breaks=seq(N1, N2, k)) : 데이터의 변수를 범위를 N1~N2, 간격을 k
만들하는 히스토그램을 만든다. > 앞에 log를 추가하면 상대적으로 정규분포 형태를 띠다.
- 0 durbinWatsonTest(데이터) : 오차의 자기상관 확인

다중회귀분석 실습####

1) 상관분석 실행###

```
corr.test(bicycle[, c(7:14)], method = "pearson",
           alpha = 0.05, use="pairwise.complete.obs")
#bicycle 데이터 프레임에서 모든 행에서 7열~14열까지
#---> 결과: 2p-value가 알파인 0.05와 비교했을 때
# 대부분의 값이 0이므로 모든 변수들 간에 상관관계와
# 상관계수 추정치가 통계적으로 유의하다. (2p-value<0.05)
# 가장 강한 양의 상관관계: temp & atemp(0.99)
```

2) 연구 가설 수립###

```
# IV: temp, atemp, humidity, windspeed, difference & DV
: total
```

#H1: temp → total(+) #temp는 total에 양의 인과관계를
미친다/ 두 변수간에는 양의 인과관계를 갖고 있다./temp가
높아질 수록 total도 높아진다.

#H2: atemp → total(+)

#H3: humidity → total(-)

#H4: windspeed → total(+)

#H5: difference → total(+)

3) 다중회귀식 수립###

```
lm1 <- lm(total ~ temp + atemp + humidity + windspeed
+ difference, data=bicycle) lm1 <- lm(종속변수 ~ 독립변수)
# yi(hat) = a + b1x1i + b2x2i + ... + b5x5i
# x1i = temp.... x5i = difference
```

4) 다중회귀분석 전제조건 확인 : plot

plot(lm1)

#residuals vs fitted: 빨간색 선이 하얀색 점선을 따라서 일직선이어야 선형성 조건 만족(그렇지 않음)
#normal Q-Q: 점들이 대각선 위에 존재하면 정규성 조건 만족(그렇지 않음)

#scale-location: 빨간색 선이 일직선이어야 등분산성과 선형성 조건 만족(그렇지 않음)

#residuals vs leverage: leverage는 사례가 다른 사례로부터 떨어진 정도로 0에 가까울 수록 바람직함. 일부 사례가 떨어져 존재하여 이상치로 볼 수도 있음. 빨간선은 수평의 일직선이 바람직함(등분산성). cook's distance는 0.5나 혹은 1을 넘으면 해당 사례가 회귀계수 추정치에 지나치게 많은 영향을 미침(그런 사례가 없음) (있으면 0.5, 1이라고 한다)
#6720, 6721, 6722, 8915, 8917, 8918을 제거하면 선형성 /정규성/등분산성이 개선됨(각 case 번호는 6728, 6729..)

```
bicycle <- bicycle %>% filter(case != 6728, case != 6729,  
case != 6730, case != 9004, case != 9006, case != 9007)  
lm1 <- lm(total ~ temp + atemp + humidity + windspeed  
+ difference, data = bicycle) # 6개 사례 제외한 bicycle로 업데이트
```

5) 정규성 조건 확인 ## 종속변수를 기준으로 check

```
ks.test(bicycle$total, pnorm,  
mean = mean(bicycle$total), sd = sd(bicycle$total))  
#Kolmogorov-Smirnov test: n이 클 때 정규성 검토. 2p  
-value = 0 < 0.05 즉 유의하기에 정규성 조건 만족 못함.  
0.2(pValue) > 0.05(alpha) 여기 정규성 안족 → 2리포, 통계적으로 검증아  
shapiro.test(bicycle$total)
```

#shapiro.test: n이 작을 때 정규성 검토(현재는 n이 10723이기에 실행하면 오류가 뜸)

```
hist(bicycle$total, breaks = seq(0, 1000, 10))  
hist(log(bicycle$total), breaks = seq(0, 10, 0.1))  
#log: 밑이 e(무리수)인 자연로그
```

6) 독립성(오차의 자기상관) 검토##

```
library(car)  
durbinWatsonTest(lm1)
```

오차는 서로 양의(0.9153073) 자기상관 존재(p-value=0 혹은 rho != 0, 즉 상관관계가 있고 독립성이지 않는다.)
#DW가 0.1693747로 2보다 많이 멀리 떨어져 있기에 오차의 자기상관이 존재한다(예기치 않은 존재) 0에 가깝기에 자기상관이 존재하지 않는다. 4에 가까우면 음의 자기상관 존재

오차의 자기상관이 존재하지 않아야 독립성 조건에 만족함

(즉 DW값이 2미만이거나 하)

〈다중회귀분석 Real 실습〉

① 다중회귀분석 수정 평균을 통한 가설 검정

• **summary (lm1)** : summary 함수에 모니터 안드 회귀식을 넣어서 회귀분석 수정치 (Estimate)와 유의치 (P-value)를 살펴보면 P-value VS O형태, R에서 나온 P-value값은 2Pvalue*2(=0.05*2)=0.1(2회비교)

② 수정된 다중회귀식 구하기

$$Y_{\text{Chat}} = 44.630 (\text{Intercept} \text{의 Estimate}) + 1.125 X_1 T + 4.129 X_2 T$$

(①에서 나온 P-value값은 2P=0.1을 비교해서 유의한 변수들의 Estimate를 이용해서 수정 다중회귀식 작성)

(Multiple R: 0.7612, adjusted R: 0.7611) 를 살펴보면 적합도, 설명력이 높은 회귀식이다)

(P-value < 2.2e-16으로 0에 가깝고 유의 즉, 모든 회귀변수가 의미적인 귀무가설 기각하고 제거도 하나의 회귀식은 의미 매김이 있는 대량가설 선택)

③ 모형 적합도 제고를 위한 다중회귀식 수정 방법

• lm1_f <- step(lm1, direction = "forward") : lm1 데이터로 'forward' 다중회귀식 추정

• lm1_b <- step(lm1, direction = "backward") : lm1 데이터로 'backward' 다중회귀식 추정

• lm1_s <- step(lm1, direction = "both") : lm1 데이터로 'forward' 'backward' 혼합

• **summary(lm1_f)** : summary 함수를 이용해서 위에 데이터 형성할 때 나온 start : AIC=k
의 값을 각각 비교해서 작은 값이나는 방법을 택함 (결과가 3가지 방법 모두 고려 일의 과정은 변수를 제거하는가 다르다)

④ 다중공선성 확인

VIF(lm1) : VIF함수를 이용해서 다중공선성을 확인

• 우선 library(corr) 먼저하기

• 이후 VIF의 값이 5.3을 넘고 그 중 제일 값이 큰 변수를 제외하고 앞선 과정 실행하기 (Attemp 변수 제외)

• lm2 <- lm(total ~ temp + humTdTh + windSpeed + difference, data = btcycle)

• **VIF(lm2)** : Attemp 변수 제외하고 만들 회귀식의 다중공선성 확인

• **summary(lm2)** : 새로운 회귀식의 표준화 회귀식 표준회귀식 (P-value, Estimate ...)

• 동일한 방법으로 temp 변수도 제외한 데이터 lm3을 만들고 VIF, summary로 정리

⑤ IV의 중요도 (표준화 회귀변수 수정치의 절대값 크기)

• lm.beta(lm1) : 표준화 회귀변수 수정치들을 절대값 기준으로 크기 비교해서 중요도 확인

• ⑥ 새로운 변수 (IV) 추가 타당성 검토 → 우선 library(lm.beta) 먼저하기

• lm4 <- lm(total ~ humTdTh + windSpeed + difference, data = btcycle) : lm4를 만듬

• **summary(lm4)** : R-squared 및 adjusted R-squared의 값을 확인해서 추가 타당성 검토

• **anova(lm3, lm4)** : R2의 변화량인 0.0405가 통계적으로 유의한지 확인

• 이후 summary를 이용해서 lm4에 대한 수정 결과 차세이 확인

• lm5 <- lm(total ~ humTdTh + windSpeed + difference + working + season, data = btcycle)

• 두 번째 추가 IV로 season 추가하기

① summary(lm5)

② ANOVA(lm4, lm5) : R^2 의 변화량인 0.0348이 통계적으로 유의한지 확인
 ④ 이후 summary를 이용해 lm5에 대한 추정값과 차이를 확인

③ VIF(lm5) : 다중공선성 확인

※ 우선 library(car) 먼저해
 $VIF^{(1)} / (2 \times DF)$ 의 값이 2보다 작으면 다중공선성은 문제없음

⑦ 새로운 독립변수 atemp 추가 확인

⑧ lm6 <- lm(total ~ humdity + windspeed + difference + working + season + atemp, data = data)

{ atemp 변수 추가}

⑨ summary(lm6)

ANOVA(lm5, lm6) : R^2 의 변화량인 0.032가 통계적으로 유의한지 확인

⑩ VIF(lm6) : 다중공선성 확인

⑪ lm_beta(lm6) : 각 독립변수의 상대적 중요도를 절대값 기준으로 확인

⑫ 가장 영향력이 높은 lm5에 기반하여 IV 속성값이 존재할 때 종속변수(예측치) 테스트

: lm5 퍼센트에 따라 부여한 속성을 넣어서 예측치 구함

→ 새로운 독립변수를 추가했을 때 모형설명력과 짜합도가 좋아졌다고 우연히 이 변수를
 추가하는 것 아니고, 상대적으로 중요도가 떨어지는 변수들에 의해 왜곡이 발생한 것 (season fall 등)

⑬ 조절효과 확인하기

⑭ b7cycle\$working <- as.numeric(b7cycle\$working) : working 변수를 수치형으로 변환

⑮ b7cycle %>% mutate(Tinter = humdity * working) : 조절효과에 쓰일 새 변수

⑯ lm7 <- lm(total ~ humdity + windspeed + difference + working + Tinter, data = b7cycle)

: Tinter를 추가한 퍼센트입니다 (lm4와 lm7 추가)

⑰ summary(lm7)

⑱ ANOVA(lm4, lm7) : R^2 의 변화량인 0.0122가 통계적으로 유의한지 확인

※ 이후 조절효과를 살펴보기 확인

<정리된 기말정리> <로짓회귀분석>
 데이터파일이용 <read_csv("파일.csv")> : 데이터 파일 불러오기
 | 우선 library(readr) 먼저하기

summary(데이터) : 성도 및 NA 존재 확인하기

str(데이터) : 해당 데이터의 성도세계 (ex) num, chr...)

desc <- describe(데이터[A:B]) : 데이터에서 A~B까지의 변수의 범위의 기술통계량 확인

desc <- desc %>% mutate(U=mean+3*sd, LL=mean-3*sd) : 이상치 기준추가

| 우선 library(dplyr) / library(psych) 먼저하기

logit1 <- glm(종속변수 ~ 독립변수, data=데이터이용, family=binomial(c)) : 로짓회귀모형
 outterTest(logit1) : 이상치 확인하기 (위에 로짓회귀모형 바탕으로) → 우선 library(car) 먼저하기
 터이터 <- 데이터%>% filter(변수 ~ 조건) : 변수가 조건에 맞는 행들만 추출해서 데이터에 할당
 | 우선 library(dplyr) 먼저하기

summary(logit1) : 로짓회귀모형에서 중요한 결과들을 설명

hostem.test(logit1\$y, logit1\$fitted.values) : logit1의 모형 적합도 확인
 | 우선 library(resourceSelection) 먼저하기

pseudoR(logit1, which=c("CoxSnell", "NagelKerke")) : logit1의 모형 적합도 확인
 | 우선 library(descTools) 먼저하기

→ 이후 logit1에 Intern 변수 추가한 logit2 만들기 → 이상치 확인 제거하고 로짓회귀모형 다시 돌립

→ summary로 각각의 그림을 비교해서 logit2가 더 적합함을 확인 후 종영

difference <- logit1\$deviance - logit2\$deviance

df <- logit1\$df.residual - logit2\$df.residual

| 1 - pchisq(difference, df)

| 나오는 값으로 적합도 확인

→ 이후 logit2 회귀모형 대상으로 hostem.test 함수와 pseudoR 함수 이용해서 모형 적합도 확인

PREDICTION <- predict(logit2, newdata=employ) : logit value로

PREDICTION <- tfekec(PREDICTION<0, 0, 1) : 위에 값을 0을 기준으로 0, 1로 변경

employ \$\$ result <- as.factor(employ \$\$ result) : result의 성도를 범주형으로 변경
 (PREDICTION과 동일하게 변경)

ConfusionMatrix(result, prediction) : 결과를 통해 hit rate 구하기 (=accuracy)

C1001 <- data.frame(TD=100, GPA=3.5, ...) ~ : 문제에 맞게 만들

Predict(logit2, C1001) : logit2를 이용해 TD, GPA에 따른 예측

| 생성한 모델에 예측을 원하는지 체크

<

[문제부]

1-계층군집분석

데이터 이름

distsy(CHRA-hr, method="sower")

o distance-hr <- distsy(CHRA-hr, method="sower")
사례간의 거리측정하여 거리 행렬 생성

L 원 I tapply(C cluster) 단자하기

CHRA-CA-hr <- hclust(dstance-hr, method="ward.D2") ; 위에서 만든
각자에 hclust 적용

o plot(CHRA-CA-hr, col="색깔이름", main="HRA") ! 텐드로 2群 만들기

o HRA-hr \$별수 <- as.numeric(HRA-hr \$별수) : 계량형 칙도로 변형

o set.seed(k) : 실행할 때마다 결과가 달라지는 것을 막기 위해 실행 (k는 아무 숫자)

o HRA-hr-NC1 <- Mclust(HRA-hr, dstance="euclidean", min.nc=1, max.nc=5, method="average") ; 최소 ~ 최대 b를 설정하고 최적의 군집개수 찾기

o HRA-hr-HCA <- cutree(HRA-CA-hr) q) : 텐드로 2群에서 9개만큼 잘라내
(군집분석행) L HRA-hr NC1에서 나온 차이o result-hr <- aggregate(HRA-hr, by=lst(cluster=HRA-hr-HCA), mean)
; 군집별 차이와 특징 5줄

→ I tapply(Mcclus) 단자하기

2-비계층군집분석

o HRA-hr.K <- scale(CHRA-hr-k) ; 변수의 표준화한 데이터 만들기

o HRA-hr-NC2 <- Nbclust(HRA-hr-k, dstance="euclidean", min.nc=1, max.nc=5, method="kmeans") ; 최적의 군집개수 구하기

o HRA-hr-kCA <- kmeans(HRA-hr-k, centers=k, nstart=25)
; 위에서 원본회원의 군집 개수 k를 적용하여 비계층적 군집분석 실행

o result-hr2 <- aggregate(HRA-hr, by=lst(cluster=HRA-hr-kCA \$), mean) ; 군집별 차이와 특징 5줄

→ 원 I tapply(Mb clust 단자)

< KNN >

o cancer_z <- as.data.frame(scale(Cancer[2:3])) ; [-1] : 행을 제외
제외하고 마지막 1열 브루들을 표준화해서 데이터 생성o Tnd <- sample(2, nrow(cancer_z), replace=T, prob=c(0.7, 0.3))
; test와 train을 구분하는 Tnd가 거시생성

o cancer_train <- cancer_z[Tnd=1,] ; [0.7, 0.3] 배율 적용

o cancer-test <- cancer_z[Tnd=2,] > train과 test 나누어서 생성

- ① $\text{grid} \leftarrow \text{expand. grid}(k=3, 10)$: K방위를 3x10 이어서 설정
 ↳ 우선 ITlibrary (caret) 먼저 하기
- ② $\text{control} \leftarrow \text{trainControl(method = "repeated CV", number = 10, repeats = 5)}$: 최종 방법 선택
 ↳ repeated CV : 각각 다른 번수 사용 "N."
 $\text{knn.train} \leftarrow \text{train}(\text{diagnosis}, \text{data} = \text{cancer.train}, \text{method} = "knn")$,
 $\text{+ Control} = \text{control}$, $\text{fune}(\text{grid} = \text{grid})$: 최종의 KNN 선정
 ↳ ITlibrary (caret)
- ③ $\text{varImp(knn.train, scale = F)}$: IV의 중요도 확인
 ↳ 앞서 scale 적용했기 때문에
- ④ $\text{pred.test} \leftarrow \text{predict}(\text{knn.train}, \text{newdata} = \text{cancer.test})$:
 생성한 knn.train을 test에 데이터에 적용해서 성능 판단
- ⑤ $\text{confusionMatrix(pred.test, cancer.test} \$ \text{diagnosis})$: 실제값 VS 예측값
 ↳ ITlibrary (caret) ↳ 예측값 ↳ 실제값
- ⑥ $\text{knn.train} \leftarrow \text{train.knn}(\text{diagnosis}, \text{data} = \text{cancer.train}, \text{K} = 25, \text{distance} = 2, \text{kernel} = \text{c}("다양한 10가지 방법 적용"))$
 ↳ 최종의 KNN과 방법 찾기 ↳ 유율리더먼트
- ⑦ $\text{knn.pred.test} \leftarrow \text{predict}(\text{knn.train}, \text{newdata} = \text{cancer.test})$
 ↳ 개선된 모델과 실제값 적용해서 비교
- ⑧ $\text{confusionMatrix(knn, pred.test, cancer.test} \$ \text{diagnosis})$
 ↳ 개선된 모델로 실제값 VS 예측값

< SVM 실습 >

- ① $\text{lsvm} \leftarrow \text{tune.svm}(\text{diagnosis}, \text{data} = \text{cancer.SVM-train}, \text{kernel} = "lsvm", \text{cost} = \text{c}(0.1, 0.25, 0.5, 0.75, 1, 2, 3, 4, 5, 7, 10))$
 ↳ 선행(lsvm) 방식으로 train 훈련 ↳ 동선(파라미터) 값 지정
 ↳ 우선 ITlibrary(e1071) 먼저 하기
- ② $\text{lsvm.test} \leftarrow \text{predict(lsvm, newdata = cancer.SVM-test)}$, $\text{newdata} = \text{cancer.SVM-test}$
 ↳ 생성한 lsvm의 최종의 모델을 test에 적용해서 성능 판단
- ③ $\text{confusionMatrix(lsvm.test, cancer.SVM-test} \$ \text{diagnosis})$
 ↳ Accuracy, kappa 같은 값을 확인 → 예측값 VS 실제값
 ↳ ITlibrary (caret) 먼저 하기
 => 나중에 3가지 방법 중 일하계 진행