Regression for 2011 US Census PUMS data

# 1. 2011 US Census PUMS 데이터

분석 목표

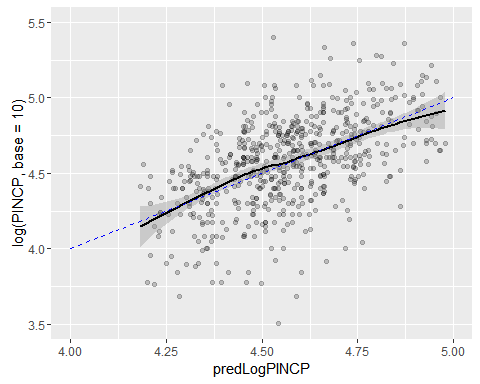
* 나이나 교육정도 등 인구통계 변수들을 이용하여 개인의 수입을 예측
* 학사학위가 수입에 미치는 영향

데이터 처리

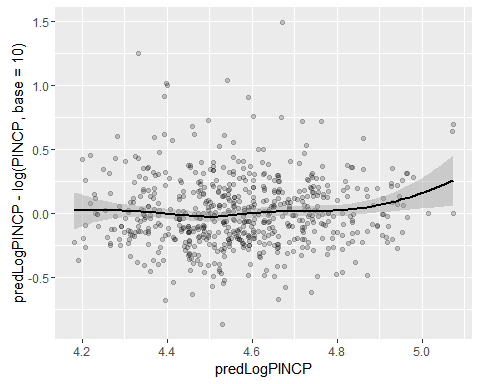
* 20~50세의 전일제 고용인으로 수입이 $1,000 ~ $250,000인 경우만 고려
* 훈련 데이터 dtrain, 시험 데이터 dtest로 분할

<https://github.com/WinVector/zmPDSwR/raw/master/PUMS/psub.RData> 에서 psub.RData를 다운 받을 수 있음

load("D:/Dropbox/PDSwR/PUMS/psub.RData")  
dtrain <- subset(psub, ORIGRANDGROUP >= 500)  
dtest <- subset(psub, ORIGRANDGROUP < 500)  
  
# COW: class of worker, SCHL: level of education, PINCP: personal income  
# M, Employee of a private for-profit, no high school diploma가 기준 범주들  
model <- lm(log(PINCP, base=10) ~ AGEP + SEX + COW + SCHL,data=dtrain)  
  
dtest$predLogPINCP <- predict(model,newdata=dtest)  
dtrain$predLogPINCP <- predict(model,newdata=dtrain)  
  
  
# 로그 수입과 예측값   
library(ggplot2)  
ggplot(data=dtest, aes(x=predLogPINCP, y=log(PINCP,base=10))) +  
 geom\_point(alpha=0.2, color="black") + #alpha=0.2는 투명도..  
 geom\_smooth(aes(x=predLogPINCP, #geom\_smooth 는 추세선(비선형)을그리라는거  
 y=log(PINCP,base=10)), color="black") +  
 geom\_line(aes(x=log(PINCP, base=10), #geom\_line은 선형회귀선(회색)을 그린거.  
 y=log(PINCP,base=10)), color="blue", linetype=2) +  
 scale\_x\_continuous(limits=c(4,5)) +  
 scale\_y\_continuous(limits=c(3.5,5.5))



# 잔차   
ggplot(data=dtest,aes(x=predLogPINCP,  
 y=predLogPINCP-log(PINCP, base=10))) +  
 geom\_point(alpha=0.2, color="black") +  
 geom\_smooth(aes(x=predLogPINCP,  
 y=predLogPINCP-log(PINCP, base=10)),  
 color="black")



특별한 문제점이 눈에 띄이지는 않는듯

# R^2  
rsq <- function(y,f) { 1 - sum((y-f)^2)/sum((y-mean(y))^2) }  
rsq(log(dtrain$PINCP,base=10), predict(model,newdata=dtrain))

## [1] 0.3382568

rsq(log(dtest$PINCP,base=10), predict(model,newdata=dtest))

## [1] 0.2605496

# sqrt(MSE)  
rmse <- function(y, f) { sqrt(mean( (y-f)^2 )) }  
rmse(log(dtrain$PINCP, base=10),predict(model, newdata=dtrain))

## [1] 0.2651856

rmse(log(dtest$PINCP, base=10),predict(model, newdata=dtest))

## [1] 0.2752171

# 훈련과 시험데이터에서 잔차  
summary(log(dtrain$PINCP, base=10) - predict(model, newdata=dtrain))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.29200 -0.14150 0.02458 0.00000 0.17630 0.62530

summary(log(dtest$PINCP, base=10) - predict(model, newdata=dtest))

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.494000 -0.165300 0.018920 -0.004637 0.175500 0.868100

## 연습문제

적절한 변수 선택을 적용하고 최종 적합 모형에 대하여 해석해보시오.

# 2. CDC 2010 출생률 데이터

신생아는 생후 1분과 5분후에 Apgar 테스트(0~10점)를 통해 즉시 응급처치 혹은 치료를 받아야 하는지 결정하며, 7점 이하면 주의를 요함(드문 경우)

* 목표: 미리 리스크가 아기를 선별하여 리소스를 효율적으로 잘 배치하고자 함
* 원데이터는 50개 주에서 모든 신생아에 대한 기록으로 엄마, 아빠, 분만에 대한 정보를 포함하며, 26,000건의 표본을 뽑은 데이터를 이용하고자 함
* <https://github.com/WinVector/zmPDSwR/raw/master/CDC> 에서 NatalRiskData.rData를 다운 받을 수 있음

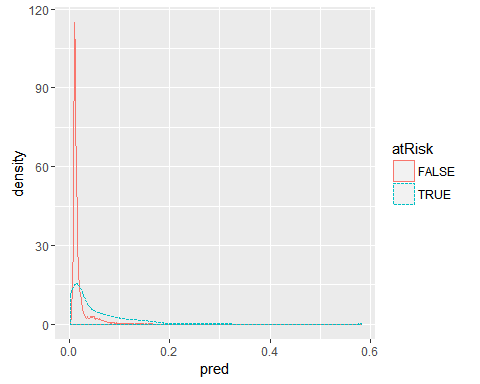
변수 설명

* atRisk: Apgar 점수가 < 7 이면 TRUE
* PWGT: 엄마의 체중
* UPREVIS: 출생전 내원 횟수
* CIG\_REC: 흡연자면 TRUE
* GESTREC3: 37주 미만(조산) 여부
* DPLURAL: single / twin / triplet+
* ULD\_MECO: 양수 찌꺼기의 오염이 보통/심한 경우 TRUE
* ULD\_PRECIP: 3시간 미만의 짧은 출산인 경우 TRUE
* ULD\_BREECH: 골반위 분만(역위)이면 TRUE
* URF\_DIAB: 엄마가 당뇨병이면 TRUE
* URF\_CHYPER: 엄마가 만성 고혈압이면 TRUE
* URF\_PHYPER: 엄마가 임신으로 인한 고혈압이면 TRUE
* URF\_ECLAM: 엄마에게 자간(임신중독증의 일종)이 있었으면 TRUE

# 데이터 읽고 분할하기  
load("D:/Dropbox/PDSwR/CDC/NatalRiskData.rData")  
train <- sdata[sdata$ORIGRANDGROUP<=5,]  
test <- sdata[sdata$ORIGRANDGROUP>5,]  
  
# 모형 공식 만들기  
complications <- c("ULD\_MECO","ULD\_PRECIP","ULD\_BREECH")  
 riskfactors <- c("URF\_DIAB", "URF\_CHYPER", "URF\_PHYPER",  
 "URF\_ECLAM")  
y <- "atRisk"  
x <- c("PWGT",  
 "UPREVIS",  
 "CIG\_REC",  
 "GESTREC3",  
 "DPLURAL",  
 complications,  
 riskfactors)  
fmla <- paste(y, paste(x, collapse="+"), sep="~")  
print(fmla)

## [1] "atRisk~PWGT+UPREVIS+CIG\_REC+GESTREC3+DPLURAL+ULD\_MECO+ULD\_PRECIP+ULD\_BREECH+URF\_DIAB+URF\_CHYPER+URF\_PHYPER+URF\_ECLAM"

# 로지스틱 회귀 적합 및 예측  
model <- glm(fmla, data=train, family=binomial(link="logit"))  
train$pred <- predict(model, newdata=train, type="response")  
test$pred <- predict(model, newdata=test, type="response")  
  
# 훈련데이터에서 그룹별 스코어값 분포  
library('ggplot2')  
ggplot(train, aes(x=pred, color=atRisk, linetype=atRisk)) +  
 geom\_density()

 두 그룹의 분리가 잘 안됨. 실제 위험한 아기는 rare임. 따라서 분류기준에 따라 살펴볼 필요가 있음

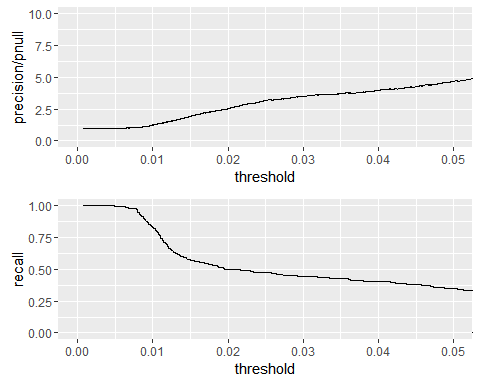
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(grid)   
  
predObj <- prediction(train$pred, train$atRisk)   
precObj <- performance(predObj, measure="prec")   
recObj <- performance(predObj, measure="rec")   
  
precision <- (precObj@y.values)[[1]]   
prec.x <- (precObj@x.values)[[1]] # 분류기준값   
recall <- (recObj@y.values)[[1]]  
  
rocFrame <- data.frame(threshold=prec.x, precision=precision,  
 recall=recall)   
  
# 여러 그림을 한 페이지 쌓아서 출력  
nplot <- function(plist) {   
 n <- length(plist)  
 grid.newpage()  
 pushViewport(viewport(layout=grid.layout(n,1)))  
 vplayout=function(x,y) {viewport(layout.pos.row=x, layout.pos.col=y)}  
 for(i in 1:n) {  
 print(plist[[i]], vp=vplayout(i,1))  
 }  
}  
  
# 훈련데이터에서 위험한 출산의 비율 계산   
pnull <- mean(as.numeric(train$atRisk))   
  
# 분류기준값의 함수로서 enrichment, recall 출력  
p1 <- ggplot(rocFrame, aes(x=threshold)) +   
 geom\_line(aes(y=precision/pnull)) +  
 coord\_cartesian(xlim = c(0,0.05), ylim=c(0,10) )  
  
p2 <- ggplot(rocFrame, aes(x=threshold)) +   
 geom\_line(aes(y=recall)) +  
 coord\_cartesian(xlim = c(0,0.05) )  
  
nplot(list(p1, p2))



분류기준이 0.02인 경우 confusion matrix, precision, recall 등

ctab.test <- table(pred=test$pred>0.02, atRisk=test$atRisk) # Note: 1   
ctab.test # Note: 2

## atRisk  
## pred FALSE TRUE  
## FALSE 9487 93  
## TRUE 2405 116

precision <- ctab.test[2,2]/sum(ctab.test[2,])  
precision

## [1] 0.04601349

recall <- ctab.test[2,2]/sum(ctab.test[,2])  
recall

## [1] 0.5550239

enrich <- precision/mean(as.numeric(test$atRisk))  
enrich

## [1] 2.664159

시험데이터에서 TP의 55.5% 찾아줌, 전체 평균에 비해 2.66배 높은 비율로...

# 모형의 계수  
coefficients(model)

## (Intercept) PWGT UPREVIS   
## -4.41218940 0.00376166 -0.06328943   
## CIG\_RECTRUE GESTREC3< 37 weeks DPLURALtriplet or higher   
## 0.31316930 1.54518311 1.39419294   
## DPLURALtwin ULD\_MECOTRUE ULD\_PRECIPTRUE   
## 0.31231871 0.81842627 0.19172008   
## ULD\_BREECHTRUE URF\_DIABTRUE URF\_CHYPERTRUE   
## 0.74923672 -0.34646672 0.56002503   
## URF\_PHYPERTRUE URF\_ECLAMTRUE   
## 0.16159872 0.49806435

# 모형 요약  
summary(model)

##   
## Call:  
## glm(formula = fmla, family = binomial(link = "logit"), data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.9732 -0.1818 -0.1511 -0.1358 3.2641   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.412189 0.289352 -15.249 < 2e-16 \*\*\*  
## PWGT 0.003762 0.001487 2.530 0.011417 \*   
## UPREVIS -0.063289 0.015252 -4.150 3.33e-05 \*\*\*  
## CIG\_RECTRUE 0.313169 0.187230 1.673 0.094398 .   
## GESTREC3< 37 weeks 1.545183 0.140795 10.975 < 2e-16 \*\*\*  
## DPLURALtriplet or higher 1.394193 0.498866 2.795 0.005194 \*\*   
## DPLURALtwin 0.312319 0.241088 1.295 0.195163   
## ULD\_MECOTRUE 0.818426 0.235798 3.471 0.000519 \*\*\*  
## ULD\_PRECIPTRUE 0.191720 0.357680 0.536 0.591951   
## ULD\_BREECHTRUE 0.749237 0.178129 4.206 2.60e-05 \*\*\*  
## URF\_DIABTRUE -0.346467 0.287514 -1.205 0.228187   
## URF\_CHYPERTRUE 0.560025 0.389678 1.437 0.150676   
## URF\_PHYPERTRUE 0.161599 0.250003 0.646 0.518029   
## URF\_ECLAMTRUE 0.498064 0.776948 0.641 0.521489   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2698.7 on 14211 degrees of freedom  
## Residual deviance: 2463.0 on 14198 degrees of freedom  
## AIC: 2491  
##   
## Number of Fisher Scoring iterations: 7

## 연습문제

적절한 변수 선택을 적용하고 최종 적합 모형에 대하여 해석해보시오.