> # Insurance data: The data is on HuskyCT

> # create a categorical variable lowcharge which equals 1 if insurance$charges < 7000 and equals 0 otherwise

> # run the logit regression of this on age, sex, bmi, smoker, region

> # split the data by choosing 1000 observations for training and by using the other observations for testing

> # assess the accuracy of this model

>

> setwd("D:/Documents (Louis Booth)/R/Big Data")

> insurance = read.csv("insurance.csv", header=TRUE)

> set.seed(80085)

> insurance$lowcharge <- rep(0, dim(insurance)[1])

> insurance$lowcharge[insurance$charges<7000] <- 1

>

> rand <- sample(nrow(insurance), size=1000, replace=F)

> train <- insurance[rand,]

> test <- insurance[-rand,]

>

> glm.fit <- glm(lowcharge~age+sex+bmi+smoker+region, family=binomial, data=train)

> glm.probs <- predict(glm.fit, test, type="response")

>

> glm.pred <- rep(0, dim(test)[1])

> glm.pred[glm.probs>0.5] <- 1

> table(glm.pred)

glm.pred

0 1

224 114

> table(glm.pred, test$lowcharge)

glm.pred 0 1

0 206 18

1 9 105

>

> mean(glm.pred == test$lowcharge)

[1] 0.9201183