> setwd("D:/Documents (Louis Booth)/R/Term Project")

> ins <- read.csv("HealthExpend.csv")

> head(ins)

AGE ANYLIMIT COLLEGE HIGHSCH GENDER MNHPOOR insure USC UNEMPLOY MANAGEDCARE famsize COUNTIP EXPENDIP

1 30 0 0 0 0 0 0 0 0 0 3 0 0.00

2 56 1 0 1 0 0 1 1 1 1 3 0 0.00

3 55 1 1 0 0 0 1 1 0 0 2 2 16121.45

4 47 0 1 0 1 0 1 1 0 0 2 0 0.00

5 50 0 1 0 1 0 1 1 1 1 1 0 0.00

6 45 0 1 0 1 0 1 1 1 1 4 0 0.00

COUNTOP EXPENDOP RACE RACE1 REGION REGION1 EDUC EDUC1 MARISTAT MARISTAT1 INCOME INCOME1

1 0 0.00 WHITE 4 MIDWEST 2 LHIGHSC 0 MARRIED 1 MINCOME 3

2 5 2384.56 BLACK 2 SOUTH 3 HIGHSCH 1 MARRIED 1 MINCOME 3

3 42 29729.56 WHITE 4 MIDWEST 2 COLLEGE 2 MARRIED 1 HINCOME 4

4 4 110.00 BLACK 2 NORTHEAST 1 COLLEGE 2 MARRIED 1 HINCOME 4

5 43 3298.95 WHITE 4 WEST 0 COLLEGE 2 DIVSEP 3 LINCOME 2

6 5 1001.43 WHITE 4 MIDWEST 2 COLLEGE 2 MARRIED 1 MINCOME 3

PHSTAT PHSTAT1 INDUSCLASS

1 EXCE 0 TRANSINFO

2 GOOD 2

3 EXCE 0 NATRESOURCE

4 FAIR 3

5 GOOD 2

6 VGOO 1

> str(ins)

'data.frame': 2000 obs. of 28 variables:

$ AGE : int 30 56 55 47 50 45 25 61 50 19 ...

$ ANYLIMIT : int 0 1 1 0 0 0 0 0 0 0 ...

$ COLLEGE : int 0 0 1 1 1 1 0 0 1 0 ...

$ HIGHSCH : int 0 1 0 0 0 0 1 1 0 0 ...

$ GENDER : int 0 0 0 1 1 1 0 1 0 1 ...

$ MNHPOOR : int 0 0 0 0 0 0 0 0 0 0 ...

$ insure : int 0 1 1 1 1 1 1 0 1 1 ...

$ USC : int 0 1 1 1 1 1 0 1 1 0 ...

$ UNEMPLOY : int 0 1 0 0 1 1 0 0 0 0 ...

$ MANAGEDCARE: int 0 1 0 0 1 1 1 0 0 1 ...

$ famsize : int 3 3 2 2 1 4 1 2 6 6 ...

$ COUNTIP : int 0 0 2 0 0 0 0 0 0 0 ...

$ EXPENDIP : num 0 0 16121 0 0 ...

$ COUNTOP : int 0 5 42 4 43 5 0 4 0 24 ...

$ EXPENDOP : num 0 2385 29730 110 3299 ...

$ RACE : Factor w/ 5 levels "ASIAN","BLACK",..: 5 2 5 2 5 5 5 5 5 5 ...

$ RACE1 : int 4 2 4 2 4 4 4 4 4 4 ...

$ REGION : Factor w/ 4 levels "MIDWEST","NORTHEAST",..: 1 3 1 2 4 1 4 1 4 4 ...

$ REGION1 : int 2 3 2 1 0 2 0 2 0 0 ...

$ EDUC : Factor w/ 3 levels "COLLEGE","HIGHSCH",..: 3 2 1 1 1 1 2 2 1 3 ...

$ EDUC1 : int 0 1 2 2 2 2 1 1 2 0 ...

$ MARISTAT : Factor w/ 4 levels "DIVSEP","MARRIED",..: 2 2 2 2 1 2 3 2 2 3 ...

$ MARISTAT1 : int 1 1 1 1 3 1 0 1 1 0 ...

$ INCOME : Factor w/ 5 levels "HINCOME","LINCOME",..: 3 3 1 1 2 3 2 1 3 3 ...

$ INCOME1 : int 3 3 4 4 2 3 2 4 3 3 ...

$ PHSTAT : Factor w/ 5 levels "EXCE","FAIR",..: 1 3 1 2 3 5 5 5 5 1 ...

$ PHSTAT1 : int 0 2 0 3 2 1 1 1 1 0 ...

$ INDUSCLASS : Factor w/ 12 levels "","FINANCE","LEISURE",..: 12 1 7 1 1 1 1 11 10 1 ...

> names <- c("GENDER","insure", "famsize", "UNEMPLOY", "EDUC1", "MARISTAT1", "INCOME1", "PHSTAT1")

> cols <- which(names(ins) %in% names)

> ins1 <- ins[,cols]

> str(ins1)

'data.frame': 2000 obs. of 8 variables:

$ GENDER : int 0 0 0 1 1 1 0 1 0 1 ...

$ insure : int 0 1 1 1 1 1 1 0 1 1 ...

$ UNEMPLOY : int 0 1 0 0 1 1 0 0 0 0 ...

$ famsize : int 3 3 2 2 1 4 1 2 6 6 ...

$ EDUC1 : int 0 1 2 2 2 2 1 1 2 0 ...

$ MARISTAT1: int 1 1 1 1 3 1 0 1 1 0 ...

$ INCOME1 : int 3 3 4 4 2 3 2 4 3 3 ...

$ PHSTAT1 : int 0 2 0 3 2 1 1 1 1 0 ...

> summary(ins1)

GENDER insure UNEMPLOY famsize EDUC1 MARISTAT1

Min. :0.000 Min. :0.0000 Min. :0.0000 Min. : 1.000 Min. :0.000 Min. :0.000

1st Qu.:0.000 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 2.000 1st Qu.:0.000 1st Qu.:0.000

Median :1.000 Median :1.0000 Median :0.0000 Median : 3.000 Median :1.000 Median :1.000

Mean :0.527 Mean :0.7775 Mean :0.2265 Mean : 3.333 Mean :0.977 Mean :0.982

3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.: 4.000 3rd Qu.:2.000 3rd Qu.:1.000

Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :12.000 Max. :2.000 Max. :3.000

INCOME1 PHSTAT1

Min. :0.000 Min. :0.000

1st Qu.:2.000 1st Qu.:0.000

Median :3.000 Median :1.000

Mean :2.534 Mean :1.355

3rd Qu.:4.000 3rd Qu.:2.000

Max. :4.000 Max. :4.000

>

> cat("Data has", nrow(ins1), "rows and", ncol(ins1), "columns. \n")

Data has 2000 rows and 8 columns.

> sum(is.na(ins1)) / (nrow(ins1) \* ncol(ins1))

[1] 0

> apply(ins1, 2, function(x) { sum(is.na(x)) })

GENDER insure UNEMPLOY famsize EDUC1 MARISTAT1 INCOME1 PHSTAT1

0 0 0 0 0 0 0 0

> train.na.per.response <- sapply(sort(unique(ins1$insure)), function(x) { apply(ins1[ins1$insure == x,], 2, function(y) { sum(is.na(y))})})

> train.na.per.response

[,1] [,2]

GENDER 0 0

insure 0 0

UNEMPLOY 0 0

famsize 0 0

EDUC1 0 0

MARISTAT1 0 0

INCOME1 0 0

PHSTAT1 0 0

> cat("Data set - Number of duplicated rows:", nrow(ins1) - nrow(unique(ins1)), "\n")

Data set - Number of duplicated rows: 730

> ##ins1 <- unique(ins1)

>

> ##install.packages("ggplot2")

> ##install.packages("gridExtra")

> library(ggplot2)

> library(gridExtra)

> plotHist <- function(data.in, i) {

+ data <- data.frame(x=data.in[,i])

+ p <- ggplot(data=data, aes(x=factor(x))) + stat\_count() + xlab(colnames(data.in)[i]) + theme\_light() + theme(axis.text.x=element\_text(size=8))

+ return (p)

+ }

> doPlots <- function(data.in, fun, ii, ncol=3) {

+ pp <- list()

+ for (i in ii) {

+ p <- fun(data.in=data.in, i=i)

+ pp <- c(pp, list(p))

+ }

+ do.call("grid.arrange", c(pp, ncol=ncol))

+ }

>

> pdf(file="data exploration 3.pdf")

Error in pdf(file = "data exploration 3.pdf") :

cannot open file 'data exploration 3.pdf'

>

> # Plot variables

> doPlots(data.in=ins1, fun=plotHist, ii=1:4, ncol=2)

> doPlots(data.in=ins1, fun=plotHist, ii=5:8, ncol=2)

>

> # Correlations between variables

> cor.ins <- cor(ins1)

> round(cor.ins,3)

GENDER insure UNEMPLOY famsize EDUC1 MARISTAT1 INCOME1 PHSTAT1

GENDER 1.000 0.061 0.199 -0.016 -0.014 0.100 -0.104 0.084

insure 0.061 1.000 -0.052 -0.162 0.282 0.050 0.310 -0.070

UNEMPLOY 0.199 -0.052 1.000 0.045 -0.220 0.017 -0.324 0.226

famsize -0.016 -0.162 0.045 1.000 -0.202 -0.121 -0.161 -0.039

EDUC1 -0.014 0.282 -0.220 -0.202 1.000 0.032 0.430 -0.236

MARISTAT1 0.100 0.050 0.017 -0.121 0.032 1.000 -0.012 0.114

INCOME1 -0.104 0.310 -0.324 -0.161 0.430 -0.012 1.000 -0.259

PHSTAT1 0.084 -0.070 0.226 -0.039 -0.236 0.114 -0.259 1.000

>

> # Visualize relationships between variables

> ##pairs(cor.train[,1:ncol(cor.train)])

> ##install.packages("GGally")

> library(GGally)

> ggpairs(ins1)

>

> # Principal Component Analysis

>

> pc.ins <- prcomp(ins1, scale.=TRUE)

> pc.ins$rotation

PC1 PC2 PC3 PC4 PC5 PC6 PC7

GENDER 0.132769379 -0.468792847 -0.57701105 0.341844832 -0.32797494 0.44011232 -0.084934322

insure -0.350798266 -0.359950054 -0.28741286 -0.242566413 0.58540971 -0.08088639 -0.448107217

UNEMPLOY 0.384397824 -0.307284694 -0.33458233 -0.270287732 0.02852808 -0.63130008 0.313232174

famsize 0.226918488 0.436373331 -0.36680991 0.436038516 0.60172959 0.07756870 0.254635969

EDUC1 -0.500688742 -0.145931283 -0.11581210 0.029032615 -0.10079181 -0.04992772 0.693006656

MARISTAT1 0.004403735 -0.476129122 0.46996773 0.662748809 0.18999122 -0.27377566 -0.007550832

INCOME1 -0.540537058 -0.002951109 -0.03320789 -0.009503482 0.13002583 0.13483891 0.175179418

PHSTAT1 0.342301419 -0.343249282 0.32026922 -0.347777814 0.35167200 0.54733392 0.343553697

PC8

GENDER 0.06576976

insure -0.23639946

UNEMPLOY 0.27449152

famsize -0.02173335

EDUC1 -0.46994629

MARISTAT1 0.04548875

INCOME1 0.80052597

PHSTAT1 0.01453059

> summary(pc.ins)

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8

Standard deviation 1.4445 1.1618 0.9998 0.9424 0.88928 0.84408 0.79810 0.73189

Proportion of Variance 0.2608 0.1687 0.1250 0.1110 0.09885 0.08906 0.07962 0.06696

Cumulative Proportion 0.2608 0.4295 0.5545 0.6655 0.76436 0.85342 0.93304 1.00000

> par(mfrow=c(1,1))

> plot(pc.ins, type="l")

> biplot(pc.ins, xlim=c(-.07,.07), ylim=c(-.07,.07))

>

> dev.off()

null device

1

>

> # frequency tables of predictors based on level of insure

> xtabs(~ insure + GENDER, ins1)

GENDER

insure 0 1

0 236 209

1 710 845

> xtabs(~ insure + UNEMPLOY, ins1)

UNEMPLOY

insure 0 1

0 326 119

1 1221 334

> xtabs(~ insure + famsize, ins1)

famsize

insure 1 2 3 4 5 6 7 8 9 10 11 12

0 51 81 82 74 61 44 35 7 3 4 2 1

1 220 437 284 310 178 69 32 13 5 2 3 2

> xtabs(~ insure + EDUC1, ins1)

EDUC1

insure 0 1 2

0 227 178 40

1 363 688 504

> xtabs(~ insure + MARISTAT1, ins1)

MARISTAT1

insure 0 1 2 3

0 178 194 14 59

1 409 919 35 192

> xtabs(~ insure + INCOME1, ins1)

INCOME1

insure 0 1 2 3 4

0 140 48 107 100 50

1 199 68 208 498 582

> xtabs(~ insure + PHSTAT1, ins1)

PHSTAT1

insure 0 1 2 3 4

0 95 120 159 57 14

1 413 502 439 140 61

>

> png(file="US Regional Map.png")

>

> # Map to visualize regional insured levels

> count\_ne <- length(which(ins$REGION == "NORTHEAST" & ins$insure == "1"))

> count\_mw <- length(which(ins$REGION == "MIDWEST" & ins$insure=="1"))

> count\_w <- length(which(ins$REGION == "WEST" & ins$insure == "1"))

> count\_s <- length(which(ins$REGION == "SOUTH" & ins$insure == "1"))

> count\_totne <- length(which(ins$REGION == "NORTHEAST"))

> count\_totmw <- length(which(ins$REGION == "MIDWEST"))

> count\_totw <- length(which(ins$REGION == "WEST"))

> count\_tots <- length(which(ins$REGION == "SOUTH"))

>

> library(maps)

> map("state", interior=TRUE)

> map("state", region=c("Connecticut", "Maine", "Massachusetts", "New Hampshire", "Rhode Island", "Vermont", "New Jersey", "New York", "Pennsylvania"), fill=TRUE, col="orange", add=TRUE)

> map("state", region=c("Illinois", "Indiana", "Michigan", "Ohio", "Wisconsin", "Iowa", "Kansas", "Minnesota", "Missouri", "Nebraska", "North Dakota", "South Dakota"), fill=TRUE, col="red", add=TRUE)

> map("state", region=c("Delaware", "Florida", "Georgia", "Maryland", "North Carolina", "South Carolina", "Virginia", "District of Columbia", "West Virginia", "Alabama", "Kentucky", "Mississippi", "Tennessee", "Arkansas", "Louisiana", "Oklahoma", "Texas"), fill=TRUE, col="yellow", add=TRUE)

> map("state", region=c("Arizona", "Colorado", "Idaho", "Montana", "Nevada", "New Mexico", "Utah", "Wyoming", "Alaska", "California", "Hawaii", "Oregon", "Washington"), fill=TRUE, col="lightyellow", add=TRUE)

> legend("bottomleft", legend=c(paste("West:", round(count\_w/count\_totw\*100, 2), "%"), paste("South:", round(count\_s/count\_tots\*100, 2), "%"), paste("Northeast:", round(count\_ne/count\_totne\*100, 2), "%"), paste("Midwest:", round(count\_mw/count\_totmw\*100, 2), "%")), title="Percent Insured by Region", fill=c("lightyellow", "yellow", "orange", "red"))

>

>

> dev.off()

null device

1

>

> # Prepare data frame for logistic regression

> names2 <- c("GENDER","insure", "famsize", "UNEMPLOY", "EDUC", "MARISTAT", "INCOME", "PHSTAT")

> cols2 <- which(names(ins) %in% names2)

> ins2 <- ins[,cols2]

> str(ins2)

'data.frame': 2000 obs. of 8 variables:

$ GENDER : int 0 0 0 1 1 1 0 1 0 1 ...

$ insure : int 0 1 1 1 1 1 1 0 1 1 ...

$ UNEMPLOY: int 0 1 0 0 1 1 0 0 0 0 ...

$ famsize : int 3 3 2 2 1 4 1 2 6 6 ...

$ EDUC : Factor w/ 3 levels "COLLEGE","HIGHSCH",..: 3 2 1 1 1 1 2 2 1 3 ...

$ MARISTAT: Factor w/ 4 levels "DIVSEP","MARRIED",..: 2 2 2 2 1 2 3 2 2 3 ...

$ INCOME : Factor w/ 5 levels "HINCOME","LINCOME",..: 3 3 1 1 2 3 2 1 3 3 ...

$ PHSTAT : Factor w/ 5 levels "EXCE","FAIR",..: 1 3 1 2 3 5 5 5 5 1 ...

> summary(ins2)

GENDER insure UNEMPLOY famsize EDUC MARISTAT

Min. :0.000 Min. :0.0000 Min. :0.0000 Min. : 1.000 COLLEGE:544 DIVSEP : 251

1st Qu.:0.000 1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.: 2.000 HIGHSCH:866 MARRIED:1113

Median :1.000 Median :1.0000 Median :0.0000 Median : 3.000 LHIGHSC:590 NEVMAR : 587

Mean :0.527 Mean :0.7775 Mean :0.2265 Mean : 3.333 WIDOWED: 49

3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.: 4.000

Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :12.000

INCOME PHSTAT

HINCOME:632 EXCE:508

LINCOME:315 FAIR:197

MINCOME:598 GOOD:598

NPOOR :116 POOR: 75

POOR :339 VGOO:622

>

>

> # Logistic regression

> logistic <- glm((insure==1) ~ GENDER + famsize + UNEMPLOY + EDUC + MARISTAT + INCOME + PHSTAT, data=ins2, family="binomial")

> coef(logistic)

(Intercept) GENDER famsize UNEMPLOY EDUCHIGHSCH EDUCLHIGHSC

3.001409655 0.481575632 -0.133607374 0.219663519 -0.791252063 -1.333514436

MARISTATMARRIED MARISTATNEVMAR MARISTATWIDOWED INCOMELINCOME INCOMEMINCOME INCOMENPOOR

0.403623234 -0.139085668 -0.254351987 -1.243959324 -0.447337523 -1.614109130

INCOMEPOOR PHSTATFAIR PHSTATGOOD PHSTATPOOR PHSTATVGOO

-1.565983586 0.026828812 -0.255204133 0.624534735 -0.007816439

> summary(logistic)

Call:

glm(formula = (insure == 1) ~ GENDER + famsize + UNEMPLOY + EDUC +

MARISTAT + INCOME + PHSTAT, family = "binomial", data = ins2)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.6518 0.2466 0.4360 0.7049 1.6934

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 3.001410 0.285352 10.518 < 2e-16 \*\*\*

GENDER 0.481576 0.123481 3.900 9.62e-05 \*\*\*

famsize -0.133607 0.032322 -4.134 3.57e-05 \*\*\*

UNEMPLOY 0.219664 0.149173 1.473 0.1409

EDUCHIGHSCH -0.791252 0.193911 -4.080 4.49e-05 \*\*\*

EDUCLHIGHSC -1.333514 0.204560 -6.519 7.08e-11 \*\*\*

MARISTATMARRIED 0.403623 0.190672 2.117 0.0343 \*

MARISTATNEVMAR -0.139086 0.191277 -0.727 0.4671

MARISTATWIDOWED -0.254352 0.378763 -0.672 0.5019

INCOMELINCOME -1.243959 0.206491 -6.024 1.70e-09 \*\*\*

INCOMEMINCOME -0.447338 0.192756 -2.321 0.0203 \*

INCOMENPOOR -1.614109 0.256476 -6.293 3.11e-10 \*\*\*

INCOMEPOOR -1.565984 0.211963 -7.388 1.49e-13 \*\*\*

PHSTATFAIR 0.026829 0.218160 0.123 0.9021

PHSTATGOOD -0.255204 0.163296 -1.563 0.1181

PHSTATPOOR 0.624535 0.350842 1.780 0.0751 .

PHSTATVGOO -0.007816 0.167939 -0.047 0.9629

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2120.2 on 1999 degrees of freedom

Residual deviance: 1788.4 on 1983 degrees of freedom

AIC: 1822.4

Number of Fisher Scoring iterations: 5

> p.hat <- fitted(logistic)

> y.hat <- round(p.hat)

> table(y.hat)

y.hat

0 1

170 1830

> table(ins2$insure)

0 1

445 1555

> table(y.hat, y.true=ins2$insure)

y.true

y.hat 0 1

0 99 71

1 346 1484

>

> cat("Successful prediction percentage for insured individuals:", round((1484/1555)\*100,2), "%")

Successful prediction percentage for insured individuals: 95.43 %

> cat("Successful prediction percentage for uninsured individuals:", round((99/445)\*100,2), "%")

Successful prediction percentage for uninsured individuals: 22.25 %

> cat("Unsuccessful prediction percentage for insured individuals:", round((71/1555)\*100,2), "%")

Unsuccessful prediction percentage for insured individuals: 4.57 %

> cat("Unsuccessful prediction percentage for uninsured individuals:", round((346/445)\*100,2), "%")

Unsuccessful prediction percentage for uninsured individuals: 77.75 %