assignment4

Keiryn Hart, 300428418

29/06/2020

```
# Load the "readxl" package to read in data from an Excel file.
library(readxl)

# Read in the heart disease dataset.
hd <- read_xlsx("C:/Users/Keiryn Hart/Documents/Uni/Data 303/Heart Disease.xlsx", sheet = "Da
ta", na = "NA")
```

Question 1: a)

```
ncol(hd)
```

```
## [1] 16
```

```
k <- 16
i <- 1
fr.prop <- data.frame(c("Frequency (n)", "Proportion (p)"))
colnames(fr.prop) <- "Variable"
for(i in 1:k){
  count <- sum(is.na(hd[,i]))
  prop <- count/nrow(hd)
  results <- data.frame(c(count, prop))
  fr.prop <- cbind(fr.prop, results)
  colnames(fr.prop)[i+1] <- paste("VARIABLE", i, sep= "_")
}

fr.prop
```

Variable	VARIABL… <dbl>	VARIABL… <dbl>	VARIABLE_3 <dbl>	VARIABL… <dbl>	VARIABLE_5 <dbl>	VARIABL… <dbl>	VARIABL… <dbl>
Frequency (n)	0	0	105.00000000	0	29.000000000	53.0000	0
Proportion (p)	0	0	0.02476415	0	0.006839623	0.0125	0

2 rows | 2-9 of 17 columns

b.

```
hd.complete <- na.omit(hd)
missing <- nrow(hd) - nrow(hd.complete)
missing/nrow(hd)
```

```
## [1] 0.1372642
```

c.

```
hd.complete[, "SBP_CAT"] <- NA
a <- nrow(hd.complete)
for(i in 1:a){
  if(hd.complete$SBP[i] < 120){
    hd.complete$SBP_CAT[i] <- "normal"
  }
  else if(hd.complete$SBP[i] > 119 & hd.complete$SBP[i] < 130){
    hd.complete$SBP_CAT[i] <- "elevated"
  }
  else if(hd.complete$SBP[i] > 129 & hd.complete$SBP[i] < 140){
    hd.complete$SBP_CAT[i] <- "High Stage 1"
  }
  else if(hd.complete$SBP[i] > 139 & hd.complete$SBP[i] < 180){
    hd.complete$SBP_CAT[i] <- "High Stage 2"
  }
  else{
    hd.complete$SBP_CAT[i] <- "hypertensive crisis"
  }
}

hd.complete
```

SEX <dbl>	AGE <dbl>	EDUC <dbl>	SMOKER <dbl>	CIG <dbl>	BP_MED <dbl>	STROKE <dbl>	HYPER <dbl>	DIAB <dbl>	CHOL <dbl>
1	39	4	0	0	0	0	0	0	195
0	46	2	0	0	0	0	0	0	250
1	48	1	1	20	0	0	0	0	245
0	61	3	1	30	0	0	1	0	225
0	46	3	1	23	0	0	0	0	285
0	43	2	0	0	0	0	1	0	228
0	63	1	0	0	0	0	0	0	205
0	45	2	1	20	0	0	0	0	313
1	52	1	0	0	0	0	1	0	260
1	43	1	1	30	0	0	1	0	225

1-10 of 3,658 rows | 1-10 of 17 columns Previous **1** 2 3 4 5 6 … 366 Next

d. because it can provide meaninful qualitative differences where the cut off points for the different levels of the predictor can reflect better levels for the model predictions.

Question 2:

a.

```
logistic <-  glm(HD_RISK ~ SBP + DBP + factor(SEX) + AGE + factor(EDUC) + CIG + CHOL + BMI +
GLUC, family = binomial, data = hd.complete)
summary(logistic)
```

```
##
## Call:
## glm(formula = HD_RISK ~ SBP + DBP + factor(SEX) + AGE + factor(EDUC) +
##     CIG + CHOL + BMI + GLUC, family = binomial, data = hd.complete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9444  -0.5969  -0.4262  -0.2843   2.9063
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.962559   0.579182 -15.475  < 2e-16 ***
## SBP            0.018249   0.003484   5.238 1.63e-07 ***
## DBP           -0.002798   0.006385  -0.438   0.6612
## factor(SEX)1   0.544414   0.108953   4.997 5.83e-07 ***
## AGE            0.063433   0.006695   9.474  < 2e-16 ***
## factor(EDUC)2 -0.188485   0.123171  -1.530   0.1259
## factor(EDUC)3 -0.196924   0.149848  -1.314   0.1888
## factor(EDUC)4 -0.052129   0.164125  -0.318   0.7508
## CIG            0.019463   0.004191   4.644 3.42e-06 ***
## CHOL           0.002371   0.001127   2.105   0.0353 *
## BMI            0.006714   0.012639   0.531   0.5952
## GLUC           0.007186   0.001674   4.293 1.77e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3121.2  on 3657  degrees of freedom
## Residual deviance: 2758.8  on 3646  degrees of freedom
## AIC: 2782.8
##
## Number of Fisher Scoring iterations: 5
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

```
## Loading required package: carData
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 3.6.3
```

```
kable(vif(logistic), digits = 2, caption = "VIF Values")%>%
  kable_styling()
```

VIF Values

	GVIF	Df	GVIF^(1/(2*Df))

	GVIF	Df	GVIF^(1/(2*Df))
SBP	3.02	1	1.74
DBP	2.78	1	1.67
factor(SEX)	1.24	1	1.11
AGE	1.29	1	1.13
factor(EDUC)	1.10	3	1.02
CIG	1.24	1	1.11
CHOL	1.06	1	1.03
BMI	1.18	1	1.09
GLUC	1.02	1	1.01

none of the vif values exceed 1.74 and all are clearly far below 10 which indicates that there is little evidence relationg to mulitcoliniarity of the predictors and therefore there is no need to remove any of the predictors.

b.

regression equation:

#log($p^{/1-p}$) = HD_RISK + 0.0183*SBP - 0.00280*DBP + 0.544*factor(SEX)1 + 0.0634*AGE - 0.189*factor(SEX)2 - 0.197*factor(SEX)3 - 0.0521*factor(SEX)4 + 0.195*CIG + 0.00237*CHOL + 0.00671*BMI + 0.00719*GLUC

```
library(pander)
```

```
## Warning: package 'pander' was built under R version 3.6.3
```

```
pander(summary(logistic))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.963	0.5792	-15.47	5.157e-54
SBP	0.01825	0.003484	5.238	1.626e-07
DBP	-0.002798	0.006385	-0.4382	0.6612
factor(SEX)1	0.5444	0.109	4.997	5.83e-07
AGE	0.06343	0.006695	9.474	2.685e-21
factor(EDUC)2	-0.1885	0.1232	-1.53	0.1259
factor(EDUC)3	-0.1969	0.1498	-1.314	0.1888
factor(EDUC)4	-0.05213	0.1641	-0.3176	0.7508
CIG	0.01946	0.004191	4.644	3.424e-06
CHOL	0.002371	0.001127	2.105	0.0353

	Estimate	Std. Error	z value	Pr(>|z|)
BMI	0.006714	0.01264	0.5313	0.5952
GLUC	0.007186	0.001674	4.293	1.766e-05

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	3121 on 3657 degrees of freedom
Residual deviance:	2759 on 3646 degrees of freedom

c.

Walds tests: SBP:

Hypothesis: H0:β1 = 0 H1:β1 ≠ 0

test statistic: 5.238

p-value: 1.63e-07

given the fact that the p-value for SBP is significantly smaller than any significance level in consideration we have enough evidence to suggest that B1 is significantly different from 0 and therefore SBP is statistically significant as a predictor of a 10 year risk of Coronary Heart disease.

DBP:

Hypothesis: H0:β2 = 0 H1:β2 ≠ 0

Test statistic: -0.438

p-value: 0.6612

Considering that the p-value for DBP is quite large and it exceeds any significance level, a conclusion can be formed that B2 is not significantly different from 0 which means that the effect of DBP does not appear to be a statistically significant for 10 year risk of coronary heart disease.

d. an increase of 1mm Hg in SBP is associated with an estimated multiplicative change of exp(0.018249) as shown below. The resulting figure corresponds to a roughly 2% increase in the odds of having a 10 year risk of future coronary heart disease.

```
exp(0.018249)
```

```
## [1] 1.018417
```

```
kable(exp(confint.default(logistic, parm = "SBP")), digits = 3)%>%
  kable_styling()
```

	2.5 %	97.5 %
SBP	1.011	1.025

e.

```
hd.complete$SBP_CAT <- factor(hd.complete$SBP_CAT, levels = c("normal", "elevated", "High Sta
ge 1", "High Stage 2", "hypertensive crisis"))
logistic2 <- glm(HD_RISK ~ factor(SBP_CAT) + DBP + factor(SEX) + AGE + factor(EDUC) + CIG + C
HOL + BMI + GLUC, family = binomial, data = hd.complete)
summary(logistic2)
```

```
##
## Call:
## glm(formula = HD_RISK ~ factor(SBP_CAT) + DBP + factor(SEX) +
##     AGE + factor(EDUC) + CIG + CHOL + BMI + GLUC, family = binomial,
##     data = hd.complete)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7466  -0.6025  -0.4283  -0.2843   2.8458
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -7.790357   0.696157 -11.191  < 2e-16
## factor(SBP_CAT)elevated          0.234938   0.161846   1.452 0.146609
## factor(SBP_CAT)High Stage 1      0.189517   0.177903   1.065 0.286746
## factor(SBP_CAT)High Stage 2      0.592431   0.184339   3.214 0.001310
## factor(SBP_CAT)hypertensive crisis  1.155536   0.298192   3.875 0.000107
## DBP                              0.006231   0.005864   1.063 0.287955
## factor(SEX)1                     0.517430   0.108387   4.774 1.81e-06
## AGE                              0.067004   0.006634  10.100  < 2e-16
## factor(EDUC)2                   -0.190932   0.122913  -1.553 0.120329
## factor(EDUC)3                   -0.215698   0.149497  -1.443 0.149068
## factor(EDUC)4                   -0.061899   0.164188  -0.377 0.706173
## CIG                              0.019659   0.004184   4.699 2.62e-06
## CHOL                             0.002539   0.001126   2.256 0.024097
## BMI                              0.006294   0.012617   0.499 0.617921
## GLUC                             0.007432   0.001667   4.457 8.29e-06
##
## (Intercept)                     ***
## factor(SBP_CAT)elevated
## factor(SBP_CAT)High Stage 1
## factor(SBP_CAT)High Stage 2     **
## factor(SBP_CAT)hypertensive crisis ***
## DBP
## factor(SEX)1                    ***
## AGE                             ***
## factor(EDUC)2
## factor(EDUC)3
## factor(EDUC)4
## CIG                             ***
## CHOL                            *
## BMI
## GLUC                            ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3121.2  on 3657  degrees of freedom
## Residual deviance: 2768.2  on 3643  degrees of freedom
## AIC: 2798.2
##
## Number of Fisher Scoring iterations: 5
```

ii.

```
elevated <- exp(0.234938)
High1 <- exp(0.189517)
High2 <- exp(0.592431)
hypertensive <- exp(1.155536)

elevated
```

```
## [1] 1.26483
```

```
High1
```

```
## [1] 1.208666
```

```
High2
```

```
## [1] 1.808379
```

```
hypertensive
```

```
## [1] 3.175725
```

SBP_CAT:

elevated:

the p-value for the elevated level is 0.146609, this is quite large and it exceeds any significance level which means we have evidence to suggest that B1 is not significantly different from 0 which indicates that an elevated SBP level is not statistically significant in predicting a 10 year risk of coronary heart disease.

effect:

for the elevated level of SBP_CAT, having Systolic blood pressure in this range is associated with an estimated multiplicative change of exp(0.234938) which corresponds to roughly a 26% increase from a normal level in the odds of having future coronary heart disease.

High stage 1:

the p-value for high stage 1 is 0.286746, this is also a large p-value and is even larger than that of the elevated category. similarly to elevated it exceeds all significance levels and therefore B1 is not significantly different from 0 and in turn is not statistically significant in predicting a 10 year risk of coronary heart disease.

effect:

for the High stage 1 level, having having Systolic blood pressure in this range is associated with an estimated multiplicative change of exp(0.189517) which corresponds to roughly a 20% increase from the normal level in the odds of having future coronary heart disease.

High stage 2:

the p-value for high stage 2 is 0.001310, this p-value is very small and is alot smaller than any significance level, from this we can confirm that there is evidence to suggest that B1 is significantly different from 0 meaning that high stage 2 of SBP_CAT is statistically significant in predicting a 10 year risk of coronary heart disease.

effect:

for the High stage 2 level, having having Systolic blood pressure in this range is associated with an estimated multiplicative change of exp(0.592431) which corresponds to roughly a 217% increase from the normal level in the odds of having future coronary heart disease.

hypertensive crisis:

the p-value for hypertensive crisis is 0.000107, again this p-value is very small and is alot smaller than any significance level, from this we can confirm that there is evidence to suggest that B1 is significantly different from 0 meaning that hypertensive crisis is statistically significant in predicting a 10 year risk of coronary heart disease.

effect: for the hypertensive crisis level, having having Systolic blood pressure in this range is associated with an estimated multiplicative change of exp(1.155536) which corresponds to roughly a 80% increase from the normal level in the odds of having future coronary heart disease. 1.155536

my results do not agree with the findings of Wu et al. (2015), this is because I have found that the risk of coronary heart disease is higher with SBP from 120-129mm Hg sitting at 26% (in respect to the reference level) than the risk of coronary heart disease with SBP from 130 - 139mm Hg sitting at 20% (in respect to the reference level). this is a contradiction to the findings of Wu but I am not entirely sure why this is.

```
kable(exp(confint.default(logistic2, parm = "SBP_CAT")), digits = 3)%>%
  kable_styling()
```

	2.5 %	97.5 %
SBP_CAT	NA	NA

f.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.6.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.6.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
pander(lrtest(logistic, logistic2))
```

Likelihood ratio test

#Df	LogLik	Df	Chisq	Pr(>Chisq)

#Df	LogLik	Df	Chisq	Pr(>Chisq)
12	-1379	NA	NA	NA
15	-1384	3	9.35	0.02498

given that the p-value for this likelihood ration test is smaller than any reasonable significance level with a value of 0.02498, we do have sufficient evidence to suggest that we would prefer the second model that uses SBP_CAT over the first model that does not.

g.

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 3.6.3
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
pander(hoslem.test(hd.complete$HD_RISK, logistic2$fitted.values, g = 10))
```

Hosmer and Lemeshow goodness of fit (GOF)
test:
`hd.complete$HD_RISK, logistic2$fitted.values`

Test statistic	df	P value
8.273	8	0.4073

```
pander(hoslem.test(hd.complete$HD_RISK, logistic2$fitted.values, g = 20))
```

Hosmer and Lemeshow goodness of fit (GOF)
test:
`hd.complete$HD_RISK, logistic2$fitted.values`

Test statistic	df	P value
17.36	18	0.4988

```
pander(hoslem.test(hd.complete$HD_RISK, logistic2$fitted.values, g = 30))
```

Hosmer and Lemeshow goodness of fit (GOF)
test:
`hd.complete$HD_RISK, logistic2$fitted.values`

Test statistic	df	P value
17.64	28	0.9349

the Hosmer-Lemeshow tests for the second logistic regression model which included SBP_CAT show that as the amount of groups increases from 10 to 20 to 30 the p-value also increases which helps to show less and less evidence against the current model being tested meaning that these values appear to support the model

being a reasonable fit for the observed data.

this test does show that in reard to the number of groups it does apeear to be better to have less groups as opposed to more.

Question 3:

a.

```
library(MASS)
forward <- stepAIC(glm(HD_RISK ~ 1, data = hd.complete, family = "binomial"), scope = list(up
per =  ~ SEX + AGE + factor(EDUC) + factor(SMOKER) + CIG + factor(BP_MED) + factor(STROKE) +
 factor(HYPER) + factor(DIAB) + CHOL + SBP + DBP + BMI + HR + GLUC, lower = ~1), direction =
"forward", trace = FALSE)
forward$anova
```

Step <fctr>	Df <dbl>	Deviance <dbl>	Resid. Df <dbl>	Resid. Dev <dbl>	AIC <dbl>
	NA	*NA*	3657	3121.187	3123.187
+ AGE	1	200.621281	3656	2920.566	2924.566
+ SBP	1	65.007004	3655	2855.559	2861.559
+ SEX	1	49.602275	3654	2805.956	2813.956
+ CIG	1	19.987492	3653	2785.969	2795.969
+ GLUC	1	19.116019	3652	2766.853	2778.853
+ CHOL	1	4.081172	3651	2762.772	2776.772
+ factor(HYPER)	1	2.985811	3650	2759.786	2775.786
+ factor(STROKE)	1	2.286714	3649	2757.499	2775.499

9 rows

```
backward <- stepAIC(glm(HD_RISK ~ AGE + SBP + SEX + CIG + GLUC + CHOL + factor(HYPER) +
    factor(STROKE), data = hd.complete, family = "binomial"), scope = list(upper =  ~ SEX + A
GE + factor(EDUC) + factor(SMOKER) + CIG + factor(BP_MED) + factor(STROKE) + factor(HYPER) +
 factor(DIAB) + CHOL + SBP + DBP + BMI + HR + GLUC, lower = ~1), direction = "backward", trac
e = FALSE)
backward$anova
```

Step <fctr>	Df <dbl>	Deviance <dbl>	Resid. Df <dbl>	Resid. Dev <dbl>	AIC <dbl>
	NA	*NA*	3649	2757.499	2775.499

1 row

b.

```
library(bestglm)
```

```
## Warning: package 'bestglm' was built under R version 3.6.3
```

```
## Loading required package: leaps
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

```
predictors <- data.frame(SEX = hd.complete$SEX, AGE = hd.complete$AGE, EDUC = hd.complete$EDU
C, SMOKER = hd.complete$SMOKER, CIG = hd.complete$CIG, BP_MED = hd.complete$BP_MED, STROKE =
hd.complete$STROKE, HYPER = hd.complete$HYPER, DIAB = hd.complete$DIAB, CHOL = hd.complete$CH
OL, SBP = hd.complete$SBP, DBP = hd.complete$DBP, BMI = hd.complete$BMI, HR = hd.complete$HR,
GLUC = hd.complete$GLUC, y = hd.complete$HD_RISK)

#best.AIC <- bestglm(Xy = predictors, family = binomial, IC = "AIC", method = "exhaustive")

#$BestModels
```

c.

my computer is unable to perform the required subset selection in part (b).