DATA 303/473 Assignment 4

Name:Nicholas Tran, ID:300296259

Due: 2 June 2022

Assignment Questions

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.0.5
```

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

```
library(boot)
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
library(MASS)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5
```

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

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
library(pander)
## Warning: package 'pander' was built under R version 4.0.5
library(bestglm)
## Warning: package 'bestglm' was built under R version 4.0.5
## Loading required package: leaps
library(foreach)
## Warning: package 'foreach' was built under R version 4.0.5
library(doParallel)
## Warning: package 'doParallel' was built under R version 4.0.5
## Loading required package: iterators
## Warning: package 'iterators' was built under R version 4.0.5
## Loading required package: parallel
library(ResourceSelection)
## Warning: package 'ResourceSelection' was built under R version 4.0.5
## ResourceSelection 0.3-5   2019-07-22
wtp <- read.csv("WTP.csv")
```

1. Data pre-processing

a.

```
wtp = wtp %>% mutate(INCOME_NONRESPONSE =
                       case_when(is.na(TOTAL_INCOME)~1,TRUE~0))

table(wtp$INCOME_NONRESPONSE)
```

```
##
##   0   1
## 338 917
```

b.

```
wtp.reduced = wtp %>% select(TOWN,
                             SEX,
                             AGE,
                             EDUC,
                             HEAD,
                             PAY_WATER,
                             ELECTRIC,
                             TIME_LENGTH,
```

```
                              INCOME_NONRESPONSE)
head(wtp.reduced)
```

```
##   TOWN SEX AGE EDUC HEAD PAY_WATER ELECTRIC TIME_LENGTH INCOME_NONRESPONSE
## 1    0   0  27    3    1         1        1          35                   1
## 2    0   0  42    4    1         1        1          45                   1
## 3    0   0  23    4    1         0        1          30                   1
## 4    0   0  39 9999    1         1        1          31                   1
## 5    0   0  27    3    1         1        1          40                   1
## 6    0   1  22    4    0         1        1          41                   1
```

c.

```
wtp.complete = wtp.reduced %>% filter(!(AGE == 9999 | EDUC == 9999))
format(round((nrow(wtp.reduced)-nrow(wtp.complete))/nrow(wtp.reduced), 3), nsmall = 2)
```

```
## [1] "0.084"
```

A proportion of 0.084 (8.4%) observations was removed from the original dataset to produce this final dataframe.

d.

```
str(wtp.complete)
```

```
## 'data.frame':    1150 obs. of  9 variables:
##  $ TOWN              : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SEX               : int  0 0 0 0 1 1 0 1 0 0 ...
##  $ AGE               : int  27 42 23 27 22 29 19 18 30 50 ...
##  $ EDUC              : int  3 4 4 3 4 4 2 2 4 5 ...
##  $ HEAD              : int  1 1 1 1 0 0 1 0 1 1 ...
##  $ PAY_WATER         : int  1 1 0 1 1 1 1 1 1 1 ...
##  $ ELECTRIC          : int  1 1 1 1 1 1 0 1 0 1 ...
##  $ TIME_LENGTH       : int  35 45 30 40 41 38 30 50 35 38 ...
##  $ INCOME_NONRESPONSE: num  1 1 1 1 1 1 0 1 0 0 ...
```

Factors:

- TOWN
- SEX
- EDUC
- HEAD
- PAY_WATER
- ELECTRIC

```
wtp.complete$TOWN <- as.factor(wtp.complete$TOWN)
wtp.complete$SEX <- as.factor(wtp.complete$SEX)
wtp.complete$EDUC <- as.factor(wtp.complete$EDUC)
wtp.complete$HEAD <- as.factor(wtp.complete$HEAD)
wtp.complete$PAY_WATER <- as.factor(wtp.complete$PAY_WATER)
wtp.complete$ELECTRIC <- as.factor(wtp.complete$ELECTRIC)

str(wtp.complete)
```

```
## 'data.frame':    1150 obs. of  9 variables:
##  $ TOWN              : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ SEX               : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 2 1 1 ...
##  $ AGE               : int  27 42 23 27 22 29 19 18 30 50 ...
##  $ EDUC              : Factor w/ 6 levels "0","1","2","3",..: 4 5 5 4 5 5 3 3 5 6 ...
##  $ HEAD              : Factor w/ 2 levels "0","1": 2 2 2 2 1 1 2 1 2 2 ...
##  $ PAY_WATER         : Factor w/ 2 levels "0","1": 2 2 1 2 2 2 2 2 2 2 ...
##  $ ELECTRIC          : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 2 1 2 ...
##  $ TIME_LENGTH       : int  35 45 30 40 41 38 30 50 35 38 ...
##  $ INCOME_NONRESPONSE: num  1 1 1 1 1 1 0 1 0 0 ...
```

2. Inferential analysis

a.

```
log.model <- glm(INCOME_NONRESPONSE ~ SEX + AGE + EDUC + PAY_WATER + ELECTRIC,
                 family = "binomial", data = wtp.complete)
```

```
pander(vif(log.model))
```

	GVIF	Df	GVIFˆ(1/(2*Df))
SEX	1.088	1	1.043
AGE	1.279	1	1.131
EDUC	1.581	5	1.047
PAY_WATER	1.047	1	1.023
ELECTRIC	1.292	1	1.136

All predictor GVIFs are well below the threshold of 10 so there's no evidence of multicolinearity between the predictor.

b.

```
pander(summary(log.model))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.349	0.4362	3.093	0.001984
SEX1	0.1774	0.1552	1.143	0.2529
AGE	-0.001464	0.005932	-0.2468	0.805
EDUC1	-0.2338	0.3329	-0.7024	0.4824
EDUC2	-0.2421	0.3568	-0.6787	0.4973
EDUC3	-0.2003	0.3507	-0.571	0.568
EDUC4	-0.1705	0.3443	-0.4951	0.6206
EDUC5	-0.6447	0.4679	-1.378	0.1682
PAY_WATER1	-0.7448	0.1622	-4.592	4.386e-06
ELECTRIC1	0.5834	0.1553	3.757	0.0001722

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1360 on 1149 degrees of freedom
Residual deviance:	1321 on 1140 degrees of freedom

Using the Wald test hypothesis:

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0$

We have evidence that, keeping all other predictors the same, `PAY_WATER` (whether the household pays for water or not) is a statistically significant predictor of `INCOME_NONRESPONSE`. The p-value for `PAY_WATER` is less than the confidence interval of $\alpha = 0.05$, thus there's strong evidence to suggest that predictors of `PAY_WATER` are both significantly different from 0 suggesting a significant relationship between whether the household pays for water and providing income on the survey.

We also have evidence that keeping all predictors the same, `ELECTRIC` is a significant predictor of `INCOME_NONRESPONSE`.

There isn't sufficient evidence to suggest any of the other predictors have significant relationships with the response `INCOME_NONRESPONSE`.

c.

$\hat{\beta}_8 \approx -0.7448 \implies exp(\hat{\beta}_8) \approx 0.475$

$\hat{\beta}_9 \approx 0.5834 \implies exp(\hat{\beta}_9) \approx 1.79$

```
pander(exp(confint.default(log.model, parm = c("PAY_WATER1", "ELECTRIC1"))))
```

	2.5 %	97.5 %
PAY_WATER1	0.3455	0.6525
ELECTRIC1	1.322	2.43

The odds of someone providing income in the survey when their household pays for water is estimated to be 0.475 (95% CI: (0.3455, 0.6525)) that of those whose households do not pay for water. given that all other predictors are kept constant. (The odds of providing income data for households that pay for water is estimated to be 0.475 times the odds of providing income data for households that do not pay for water.)

The odds of providing income on the survey for households that is connected to the electrical grid is estimated to be 1.79 (95% CI: (1.322, 2.43)) that of those whose households are not connected to the electrical grid. (The odds of providing income data for households that are connected to the electrical grid is estimated to be 1.79 times the odds of providing income data for households that are not connected to the electrical grid.)

d.

```
log.model.interactions <- glm(INCOME_NONRESPONSE ~ SEX * PAY_WATER + AGE + EDUC
                              + SEX * ELECTRIC, family = "binomial", data = wtp.complete)
pander(summary(log.model.interactions))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.286	0.4522	2.844	0.004458
SEX1	0.4797	0.373	1.286	0.1984
PAY_WATER1	-0.5266	0.1836	-2.869	0.004124
AGE	-0.0004659	0.00597	-0.07804	0.9378
EDUC1	-0.2614	0.3395	-0.7698	0.4414
EDUC2	-0.2578	0.3632	-0.7098	0.4779
EDUC3	-0.2112	0.3574	-0.591	0.5545
EDUC4	-0.1722	0.3505	-0.4912	0.6233
EDUC5	-0.6381	0.4715	-1.353	0.176
ELECTRIC1	0.3734	0.1796	2.079	0.03758
SEX1:PAY_WATER1	-0.9166	0.3967	-2.31	0.02086
SEX1:ELECTRIC1	0.7513	0.3159	2.378	0.01739

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	1360 on 1149 degrees of freedom
Residual deviance:	1311 on 1138 degrees of freedom

Likelihood ratio test hypothesis:

For interaction between `SEX` and `PAY_WATER`

$H_0 : \beta_{10} = 0$

$H_1 : \beta_{10} \neq 0$

We get a p-value of 0.02086 which is lower than our confidence interval threshold of $\alpha = 0.05$, suggesting that the interaction between `SEX` (sex of the individual) and `PAY_WATER` (whether the household pays for water) has a significant effect on `INCOME_NONRESPONSE` (whether the household provided income data).

For interaction between `SEX` and `ELECTRIC`

$H_0 : \beta_{11} = 0$

$H_1 : \beta_{11} \neq 0$

We get a p-value of 0.01739 which is lower than our confidence interval threshold of $\alpha = 0.05$, suggesting that the interaction between `SEX` (sex of the individual) and `ELECTRIC` (whether the household is connected to the electrical grid) has a significant effect on `INCOME_NONRESPONSE` (whether the household provided income data).

e.

Let:

- X_1 denote `SEX`
- X_2 denote `AGE`
- X_3 denote `EDUC`
- X_4 denote `PAY_WATER`
- X_5 denote `ELECTRIC`

Equation for model without interaction terms:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{11} + \beta_2 X_2 + \beta_3 X_{31} + \beta_4 X_{32} + \beta_5 X_{33} + \beta_6 X_{34} + \beta_7 X_{35} + \beta_8 X_{41} + \beta_9 X_{51}$

Equation for model with interaction between `SEX` and `PAY_WATER`, and `SEX` and `ELECTRIC`:

$log(\frac{p}{1-p}) = \beta_0 + \beta_1 X_{11} + \beta_2 X_2 + \beta_3 X_{31} + \beta_4 X_{32} + \beta_5 X_{33} + \beta_6 X_{34} + \beta_7 X_{35} + \beta_8 X_{41} + \beta_9 X_{51} + \beta_{10} X_{11} X_{41} + \beta_{11} X_{11} X_{51}$

i)

$H_0 : \beta_{10} = \beta_{11} = 0$

$H_1 : \beta_{10} \neq 0$ or $\beta_{11} \neq 0$

```
pander(lrtest(log.model, log.model.interactions),caption = "")
```

#Df	LogLik	Df	Chisq	Pr(>Chisq)
10	-660.4	NA	NA	NA
12	-655.5	2	9.809	0.007412

ii)

$G^2 \approx 9.809$

iii)

Asymptotic $\chi^2_{12-10} = \chi^2_2$ distribution under H_0

iv)

p-value $\approx P(\chi^2_2 > 9.809) \approx 0.007412$

v)

The p-value is considerably lower than any confidence interval threshold thus β_{10} and β_{11} are significantly different from 0 so there's strong evidence that the model with interactions provides a better fit to the data than the model without interactions.

f.

```
pander(hoslem.test(wtp.complete$INCOME_NONRESPONSE,
                   log.model.interactions$fitted.values, g = 5))
```

Table 8: Hosmer and Lemeshow goodness of fit (GOF) test: `wtp.complete$INCOME_NONRESPONSE, log.model.interactions$fitted.values`

Test statistic	df	P value
2.56	3	0.4645

```
pander(hoslem.test(wtp.complete$INCOME_NONRESPONSE,
                   log.model.interactions$fitted.values, g = 10))
```

Table 9: Hosmer and Lemeshow goodness of fit (GOF) test: `wtp.complete$INCOME_NONRESPONSE, log.model.interactions$fitted.values`

Test statistic	df	P value
7.529	8	0.4808

```
pander(hoslem.test(wtp.complete$INCOME_NONRESPONSE,
                   log.model.interactions$fitted.values, g = 15))
```

Table 10: Hosmer and Lemeshow goodness of fit (GOF) test: `wtp.complete$INCOME_NONRESPONSE, log.model.interactions$fitted.values`

Test statistic	df	P value
12.97	13	0.4498

The p-values for the Hosmer-Lemeshow tests for $g = 5, 10, 15$ are all much greater than the confidence interval of $\alpha = 0.05$ so there's evidence to suggest that the model provides a reasonable fit to the data.

3. Statistical learning

a.

```
forward.select <- stepAIC(
  glm(INCOME_NONRESPONSE ~ 1, family = "binomial", data =wtp.complete),
  scope = list(upper = ~ TOWN + SEX + AGE + EDUC + HEAD + PAY_WATER + ELECTRIC
               + TIME_LENGTH, lower = ~1),
  direction = "forward", trace = FALSE)

pander(forward.select$anova)
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1149	1360	1362
+ PAY_WATER	1	19.11	1148	1341	1345
+ ELECTRIC	1	15.94	1147	1325	1331
+ HEAD	1	2.687	1146	1322	1330

```
backward.select <- stepAIC(
  glm(INCOME_NONRESPONSE ~ TOWN + SEX + AGE + EDUC + HEAD + PAY_WATER + ELECTRIC
      + TIME_LENGTH,
      family = "binomial",
      data = wtp.complete),
  scope = list(upper = ~ TOWN + SEX + AGE + EDUC + HEAD + PAY_WATER + ELECTRIC
               + TIME_LENGTH, lower = ~1),
  direction = "backward", trace = FALSE)

pander(backward.select$anova)
```

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	1136	1319	1347
- EDUC	5	2.362	1141	1321	1339
- TOWN	2	0.8951	1143	1322	1336
- AGE	1	0.005263	1144	1322	1334
- SEX	1	0.1155	1145	1322	1332
- TIME_LENGTH	1	0.3201	1146	1322	1330

The optimal model given by forward selection is: `INCOME_NONRESPONSE ~ HEAD + PAY_WATER + ELECTRIC`

The optimal model given by backward selection is: `INCOME_NONRESPONSE ~ HEAD + PAY_WATER + ELECTRIC`

The models given by the forward selection technique and the model given by the backward selection technique is the same model the model with only `HEAD`, `PAY_WATER` and `ELECTRIC` as predictors.

The two techniques can often yield different models due to the differences in algorithms. In forward selection we start with the null model (the model with no predictors) and continue adding predictors until a stopping criteria is reached, backward selection on the other hand starts with the full model (the model with all predictors) and start removing predictors until the stopping criteria is reached. It's possible (and usual) for the algorithms to stop at different points yielding different models.

b.

```
predictors.for.bestglm <- data.frame(TOWN = wtp.complete$TOWN,
                                     SEX = wtp.complete$SEX,
```

```
                                    AGE = wtp.complete$AGE,
                                    EDUC = wtp.complete$EDUC,
                                    HEAD = wtp.complete$HEAD,
                                    PAY_WATER = wtp.complete$PAY_WATER,
                                    ELECTRIC = wtp.complete$ELECTRIC,
                                    TIME_LENGTH = wtp.complete$TIME_LENGTH,
                                    y = wtp.complete$INCOME_NONRESPONSE)
```

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

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## Note: factors present with more than 2 levels.
```

```
pander(best.logistic.AIC$BestModels)
```

Table 13: Table continues below

TOWN	SEX	AGE	EDUC	HEAD	PAY_WATER	ELECTRIC	TIME_LENGTH
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE

Criterion
1328
1329
1329
1330
1330

```
best.logistic.BIC <- bestglm(Xy = predictors.for.bestglm, family = binomial, IC = "BIC",
                             method = "exhaustive")
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
## Note: factors present with more than 2 levels.
```

```
pander(best.logistic.BIC$BestModels)
```

Table 15: Table continues below

TOWN	SEX	AGE	EDUC	HEAD	PAY_WATER	ELECTRIC	TIME_LENGTH
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE

Criterion
1339
1343
1344
1346
1346

The optimal model given by AIC is: `INCOME_NONRESPONSE ~ HEAD + PAY_WATER + ELECTRIC` with an AIC of 1328

The optimal model given by BIC is: `INCOME_NONRESPONSE ~ HEAD + PAY_WATER + ELECTRIC` with a BIC of 1339

Both models given by AIC and BIC selection criteria are the same.

AIC and BIC may give different models due to how they penalise complexity. BIC penalises models with more predictors compared to AIC. So BIC and AIC may select different optimal models if adding a predictor yields marginal results.

c.

```
area.under.curve <- function(r, p = 0){
  require(ROCR)
  pred <- prediction(p, r)
  auc <- performance(pred, measure = "auc")
  auc@y.values[[1]]
}
```

```
nrep <- 50

nclust <- makeCluster(detectCores() * 0.75)
registerDoParallel(nclust)

variable.indices <- 1 : 8
all.comb <- expand.grid(as.data.frame(matrix(rep(0 : 1, length(variable.indices)),
                                             nrow = 2)))[-1, ]

AUC.parallel <- foreach(i = 1 : nrep, .combine = "rbind", .packages = "boot") %:%
foreach(j = 1 : nrow(all.comb), .combine = "c") %dopar%{
set.seed(1)
logistic.regression.model <- glm(as.formula(paste("INCOME_NONRESPONSE ~",
paste(names(wtp.complete)[variable.indices[all.comb[j,] == 1]], collapse = " + "))), data
= wtp.complete, family = "binomial")
return(cv.glm(wtp.complete, logistic.regression.model, cost = area.under.curve, K =
10)$delta[1])
}

stopCluster(nclust)
```

```
best.models.AUC <- (1 : nrow(all.comb))[apply(AUC.parallel, 2, mean)
                                        >= max(apply(AUC.parallel, 2, mean)
                                               - apply(AUC.parallel, 2, sd))]

for(i in 1 : length(best.models.AUC)){
cat(paste("Model ", i, ":\n"))
print(names(wtp.complete)[variable.indices[all.comb[best.models.AUC[i], ] == 1]])
```

```
print(apply(AUC.parallel, 2, mean)[best.models.AUC[i]])
cat("\n")
}
```

```
## Model  1 :
## [1] "HEAD"      "PAY_WATER" "ELECTRIC"
## [1] 0.6100647
```

The cross-validation routine for selecting an optimal model yields the same results as all the other model selection criteria: forward selection, backward selection, AIC and BIC. The model: `INCOME_NONRESPONSE ~ HEAD + PAY_WATER + ELECTRIC` with an AUC of 0.61.