DATA303-DATA473 Assignment 4

Due: 11:59 PM Thursday 2 June 2022

Background and Data

When data are collected through surveys involving human subjects, those who design the survey need to think carefully about factors that might influence whether respondents refuse to provide specific information or are unable to do so, a problem known as "non-response". Respondents are less likely to provide responses to questions on sensitive subjects. One such sensitive subject is income with Juster et al. (2006) estimating that roughly one-third of questions related to income result in non-response, although this is highly variable from survey to survey. In the United States, Lillard, Smith, and Welch (1986) reported non-response for total income of 2.5% in the 1940 Current Population Survey (CPS) and then a steady increase to 26.6% for the 1982 CPS. This high non-response rate has seemingly stabilized or decreased around the turn of the century with Moore, Stinson, and Welniak (2000) reporting a non-response rate for income questions of roughly 25% for the 1996 CPS and Dixon (2005) showing non-response for income questions to have dropped to 14.2% by the 2002-2003 CPS.

In the context of sub-Saharan African countries, Argent (2009) reported non-response according to a variety of income categories in the National Income Dynamics Study in South Africa with non-response rates ranging from 2.3% to 52.4%. In Mozambique, Fonseca (2014) reported a non-response rate of 39.5% for income questions for household surveys administered to 1,710 households across 68 communities as part of the WASHCost programme. Although estimates were not provided, Fonseca reported non-response rates to be even higher for income questions for similar surveys administered to households in Ghana as part of WASHCost.

Although the reasons for non-response may be varied, non-response to income-related questions is generally believed to be related to income of the respondent and so not missing at random. Lillard, Smith, and Welch (1986) found non-response to income-related questions to increase with income of the respondent. Biewen (2001), on the other hand, found non-response to income-related questions to be highest for those in the tails of the income distribution. In his examination of the German Socio-Economic Panel (GSOEP) study, Schräpler (2004) looked at refusals and responses of "don't know" to income-related questions and found that refusals were significantly higher with those reporting vocational positions classified as "high" (e.g., executives, civil servants) while responses of "don't know" were significantly higher with those reporting vocational positions classified as "low" (e.g., unskilled workers). As there is likely to be a fairly strong relationship between vocation classification and income, this result would appear to be in line with the findings of Biewen (2001). And Argent (2009) noted that there "is a general consensus that refusals to income questions are unlikely to be random with respect to income, with those of very high and very low incomes being less likely to respond," also in agreement with Biewen (2001).

In this assignment, we consider income data collected as part of a study carried out in three towns in northern Mozambique. This study sought to understand how much people would be willing to pay for water piped to their premises and factors that may influence how much they would be willing to pay. At the end of the survey, participants were asked a number of income-related questions, and our focus will be on factors that are associated with whether respondents provide a numeric value for their total income. A subset of variables collected as part of this study is contained in the dataset `WTP.csv`, and a list of the variables is presented in the table on the next page. Like many social surveys, there are a variety of special codes used in this dataset to reflect different types of missing data, and these are as follows:

Code	Description
-1	Respondent refused to answer the question.
9998	The question is not applicable to the respondent. This is most commonly due to a response to a previous question.
9999	The respondent specifies that they do not know.
NA	A response was not recorded.

Variable	Description
HH	Household identifier for a given enumeration area (1-15)
DAY	Day of interview
TOWN	Town (0 = “Nampula”, 1 = “Liupo”, 2 = “Ribaue”)
YEARS	Years household has lived in <TOWN>
SEX	Sex of the respondent (0 = “Male”, 1 = “Female”)
AGE	Age of the respondent (in years)
STATUS	Marital status of the respondent (0 = “Single”, 1 = “Married”, 2 = “Marital union”, 3 = “Divorced”, 4 = “Separated”, 5 = “Widowed”)
EDUC	Education level of the respondent (0 = “None”, 1 = “Primary of the 1^{st} degree”, 2 = “Primary of the 2^{nd} degree”, 3 = “Secondary of the 1^{st} degree”, 4 = “Secondary of the 2^{nd} degree”, 5 = “Higher level”)
DISABLED	Disability status of the respondent (0 = “None”, 1 = “Physical”, 2 = “Sight/visual”, 3 = “Other sensory”, 4 = “Mental”)
HEAD	Is the respondent the head of the household? (0 = “No”, 1 = “Yes”)
SEX_HH	Sex of the head of the household (0 = “Male”, 1 = “Female”)
AGE_HH	Age of the head of the household
STATUS_HH	Marital status of the head of the household
EDUC_HH	Education level of the head of the household
DISABLED_HH	Disability status of the head of the household
AGE_HH_SPOUSE	Age of the spouse of the head of the household
EDUC_HH_SPOUSE	Education level of the spouse of the head of the household
DISABLED_HH_SPOUSE	Disability status of the spouse of the head of the household
HH_SIZE	Number of persons regularly living in the household
N_ADULTS	Number of adults regularly living in the household
PRIMARY_WS	Primary water source used by household (1 = “Tap in the house”, 2 = “Tap in the yard”, 3 = “Public tap”, 4 = “Tap of a neighbour”, 5 = “Public borehole”, 6 = “Well”, 7 = “Protected spring”, 8 = “Unprotected spring”, 9 = “River, lake, or stream”). These are considered to be hierarchical with 1 considered to be the best type of water source and 9 the worst
SUFFICIENT_WATER	Does the household have sufficient access to water for its daily needs? (0 = “No”, 1 = “Yes”)
TOTAL_TIME	Total time required to travel to water source, queue for water and return home when collecting water.
PAY_WATER	Does the household pay for water? (0 = “No”, 1 = “Yes”)
TOTAL_COST	Average amount (in Mozambican meticals) the household spends each month on water-related costs (including the cost of water, water treatment, transportation of water to the home, etc.)
ELECTRIC	Is the household connected to the electrical grid? (0 = “No”, 1 = “Yes”)
TOTAL_INCOME	Average total monthly income of the household (in Mozambican meticals)
TIME_LENGTH	How long did the survey take to complete (in minutes)?

The data are available in the file WTP.csv, which can be read into R using the code below but with the path

```r
# Read in the Mozambican willingness to pay dataset.
wtp <- read.csv("WTP.csv")
```

Assignment Questions

1. **Data pre-processing: (8 marks)**

   a. **(2 marks) The variable TOTAL_INCOME records the numeric value for total income for respondents who could or were willing to provide this information Use this variable to add a new variable INCOME_NONRESPONSE to the data frame wtp. The new variable INCOME_NONRESPONSE should be a binary variable indicating whether the person did not provide a numeric total income data (0 = "Provided numeric income data", 1 = "Did not provide numeric income data"). Show your code to produce this new variable as well as a table of the frequency of outcomes of 0 and 1.**

```
# Create the variable "INCOME_NONRESPONSE" that reflects whether a numeric value
# was provided for total income.
wtp$INCOME_NONRESPONSE <- ifelse(test = wtp$TOTAL_INCOME %in% c(-1, 9998, 9999, NA),
                                 yes = 1, no = 0)
```

0	1
338	917

   b. **(2 marks) We will restrict our focus to a subset of demographic variables and variables that are generally associated with income (and so can be considered as proxies for income). In particular, we will consider a reduced dataset consisting only of the following nine variables:**

TOWN	SEX	AGE	EDUC	HEAD
PAY_WATER	ELECTRIC	TIME_LENGTH	INCOME_NONRESPONSE	

   Create a new data frame wtp.reduced that consists of only these variables. Show your code to produce this new data frame.

   Any of the following lines of code will reduce the dataset to the desired variables.

```
# Reduce the dataset to the variables "TOWN", "SEX", "AGE", "EDUC", "HEAD",
# "PAY_WATER", "ELECTRIC", "TIME_LENGTH", and "INCOME_NONRESPONSE".
wtp.reduced <- wtp[, c(3, 5 : 6, 8, 10, 24, 26, 28 : 29)]
wtp.reduced <- wtp[, c("TOWN", "SEX", "AGE", "EDUC", "HEAD", "PAY_WATER",
                       "ELECTRIC", "TIME_LENGTH", "INCOME_NONRESPONSE")]

library(dplyr)
wtp.reduced <- wtp.reduced %>% select(TOWN, SEX, AGE, EDUC, HEAD, PAY_WATER,
                                      ELECTRIC, TIME_LENGTH, INCOME_NONRESPONSE)
```

   c. **(2 marks) Now create a new data frame called wtp.complete, which only keeps respondents/observations from wtp.reduced that have no missing data. Show your code to produce this new data frame. (Note: Pay close attention to special codes for missing values.) In total, what proportion (to 3dp) of respondents/observations have been removed from the original dataset to produce this final data frame?**

```
# Recode values of -1, 9998, and 9999 to NA in "wtp.reduced".
wtp.reduced[wtp.reduced == -1 | wtp.reduced == 9998 | wtp.reduced == 9999] <- NA
# Restrict "wtp.reduced" to only observations that have no missing values,
# and save in a new data frame called "wtp.complete".
wtp.complete <- wtp.reduced[complete.cases(wtp.reduced), ]
```

```
# Calculate the proportion of observations that have been removed.
(nrow(wtp.reduced) - nrow(wtp.complete)) / nrow(wtp.reduced)
```

```
[1] 0.08366534
```

In total, 0.084 of all observations have been removed from the original dataset due to missing values for at least one variable from the reduced dataset.

d. **(2 marks) Which variables contained in `wtp.complete` are factors? List these variables, and show code to overwrite these variables in the data frame `wtp.complete` so that they are recognised by `R` as factors. (Note: DO NOT convert the variable `INCOME_NONRESPONSE` to a factor and overwrite the original variable.)**

The following variables are factors:

1) `TOWN`
2) `SEX`
3) `EDUC`
4) `HEAD`
5) `PAY_WATER`
6) `ELECTRIC`
7) `INCOME_NONRESPONSE`

Code to convert these to factors and overwrite the original variables is as shown below.

```
# Convert the variables of "TOWN", "SEX", "EDUC", "HEAD", "PAY_WATER",
# and "ELECTRIC" to factors.
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)
```

2. **Inferential analysis: (18 marks)**

Now we will focus on how non-response to a question asking for a numeric value for the total average monthly income of the household is related to demographic factors of the respondent and proxies for income.

a. **(3 marks) Fit a logistic regression model of income non-response (`INCOME_NONRESPONSE`) on sex of the respondent (`SEX`), their age (`AGE`), highest level of education completed (`EDUC`), whether the household pays for water (`PAY_WATER`), and whether the household is connected to the electrical grid (`ELECTRIC`). For this logistic regression model, calculate the variance inflation factors for predictors (to 3dp) to determine whether or not there is evidence of significant multicollinearity among the predictors in the model. If so, comment on which predictor(s) should be removed, and use this model for subsequent parts of this question.**

Code to fit the specified logistic regression and calculate variance inflation factors is as shown below.

```
# Fit a logistic regression of income non-response (INCOME_NONRESPONSE) on sex
# of the respondent (SEX), their age (AGE) highest level of education completed
# (EDUC), whether the household pays for water (PAY_WATER), and whether the
# household is connected to the electrical grid (ELECTRIC).
```

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

```r
# Load the "car" package to make use of the vif() function.
library(car)
# Calculate variance inflation factors for predictors.
vif(wtp.model)
```

Variance inflation factors (VIFs) are as shown in the table below. The largest VIF is approximately $1.136^2 \approx 1.292$, which is well below 10, alleviating concerns about multicollinearity of predictors.

Table 5: Variance inflation factors for predictors to be included in the logistic regression model.

	GVIF	Df	GVIF^(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

b. **(3 marks) Provide summary output for the logistic regression model specified in part (a). Explain what you can conclude based on Wald tests of coefficients. Provide evidence to support your conclusion.**

Summary output for the logistic regression model is as shown in the table below.

	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

Wald tests for the coefficients for `PAY_WATER1` and `ELECTRIC1` are statistically significant with p-values of 1.58×10^{-6} and 1.7×10^{-4}, respectively. Wald tests for all other coefficients (other than β_0, which is not of interest here) are non-significant.

This means that:

- there is a significant relationship between whether the household pays for water and whether the respondent does not report a numeric value for total income, adjusting for sex of the respondent, their age, the highest level of education they have completed, and whether the household is connected to the electrical grid; and

- there is a significant relationship between whether the household is connected to the electrical grid and whether the respondent does not report a numeric value for total income, adjusting for sex of the respondent, their age, the highest level of education they have completed, and whether the household pays for water.

c. **(3 marks) For any significant Wald tests in part (b), provide a precise interpetation of what the estimated coefficient suggests about the "effect" of the predictor on the response, and calculate a corresponding 95% confidence interval (to 3dp) for the estimated "effect".**

To interpret the coefficients for

i) whether the household pays for water (`PAY_WATER1`) and
ii) whether the household is connected to the electrical grid (`ELECTRIC1`),

we exponentiate the coefficients as well as corresponding 95% confidence intervals for the coefficients. The code below accomplishes this.

```
# Produce an estimate for the odds ratio giving the "effects" of PAY_WATER1
# and ELECTRICAL1
exp(wtp.model$coefficients[9 : 10])
# Produce a corresponding 95% confidence interval
pander(exp(confint.default(wtp.model, parm = c("PAY_WATER1", "ELECTRIC1)))
```

Interpreting these coefficients:

i) the odds of not reporting a numeric value for total income for respondents from households that pay for water is estimated to be 0.475 (95% CI: (0.346, 0.653)) times the odds of not reporting a numeric value for total income for respondents from households that do not pay for water, adjusting for sex of the respondent, their age, the highest level of education they have completed, and whether the household is connected to the electrical grid. (Equivalently, the odds of not reporting a numeric value for total income for respondents from households that do not pay for water is estimated to be 2.106 (95% CI: (1.533, 2.894)) times the odds of not reporting a numeric value for total income for respondents from households that pay for water, adjusting for sex of the respondent, their age, the highest level of education they have completed, and whether the household is connected to the electrical grid.)
ii) the odds of not reporting a numeric value for total income for respondents from households that are connected to the electrical grid is estimated to be 1.792 (95% CI: (1.322, 2.43)) times the odds of not reporting a numeric value for total income for respondents from households that are not connected to the electrical grid, adjusting for sex of the respondent, their age, the highest level of education they have completed, and whether the household pays for water.

d. **(3 marks) Fit the model considered in part (a) but additionally include interactions between**

- **sex of the respondent and whether the household pays for water and**
- **sex of the respondent and whether the household is connected to the electrical grid.**

Provide summary output for this model. For this model, explain what it means for sex of the respondent to interact with i) whether the household pays for water and ii) whether the household is connected to the electrical grid.

Summary output for the model including these interactions is as shown below. Interpreting what these interactions effects mean:

i) The difference between male and female respondents in terms of odds (or, equivalently, likelihood) of not providing a numeric value for total income changes depending on whether considering respondents from households that pay for water or considering respondents from households that do not pay for water.

ii) The difference between male and female respondents in terms of odds (or, equivalently, likelihood) of not providing a numeric value for total income changes depending on whether considering respondents from households that are connected to the electrical grid or considering respondents from households that are not connected to the electrical grid.

	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
ELECTRIC1	0.3734	0.1796	2.079	0.03758
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
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

e. **(3 marks) Perform an appropriate test to determine if the logistic regression model fit in part (d) provides a significantly better fit than the model that was fit in part (a). Be sure to write out the full form of the logistic regression models fit in parts (a) and (d), clearly explaining what variables represent, and state**

i) **the hypotheses of the test,**

ii) **the value of the test statistic,**

iii) **the distribution of the test statistic,**

iv) **the p-value of the test, and**

v) **your conclusion.**

The two logistic regression models being considered are

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{31} + \beta_4 X_{32} + \beta_5 X_{33} + \beta_6 X_{34} + \beta_7 X_{35} + \beta_8 X_4 + \beta_9 X_5$$

and

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{31} + \beta_4 X_{32} + \beta_5 X_{33} + \beta_6 X_{34} + \beta_7 X_{35} + \beta_8 X_4 + \beta_9 X_5 \\ &\quad +\beta_{10} X_1 X_4 + \beta_{11} X_1 X_5\end{aligned}$$

where

- X_1 denotes whether the respondent is male,

- X_2 denote the age of the respondent,
- X_3 denotes the highe level of education completed for the respondent (with 1 to 5 representing the levels of the same numbers presented in the dataset description),
- X_4 denotes whether the household pays for water, and
- X_5 denotes whether the household is connected to the electrical grid.

We perform a likelihood ratio test to compare the models fit in parts (a) and (d). Code to perform this is as shown below.

```
# Carry out a likelihood ratio test comparing the full model to the reduced model.
lrtest(wtp.model, wtp.model.int)
```

Table 10: Likelihood ratio test comparing the original logistic regression model (reduced model) with a logistic regression model that includes interactions between sex of the respondent and whether the household pays for water as well as sex of the respondent and whether the household is connected to the electrical grid (full model).

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

The likelihood ratio test is equivalent to testing the hypotheses

$$\begin{aligned}\mathcal{H}_0 : \beta_{10} &= \beta_{11} = 0\\ \mathcal{H}_1 : \beta_{10} &\neq 0 \text{ or } \beta_{11} \neq 0\end{aligned}$$

This test produces a test statistic of

$$G^2 \approx 9.809$$

which follows an asymptotic χ^2_2 distribution under $\mathcal{H}_0$

The p-value is approximately 0.0074, which is less than $\alpha = 0.05$. Thus, we reject the null hypothesis and conclude that inclusion of the interaction effects leads to a significant improvement in model fit.

f. **(3 marks) Finally, for the best model of the two you fit (in parts (a) and (d)), perform a Hosmer-Lemeshow test for $g = 5$, 10, and 15 groups, and comment on what these suggest about the goodness-of-fit of this model to the income non-response data.**

Using the model fit in part (d), which includes interaction effects, we get the results for Hosmer-Lemeshow tests based on $g = 5$, 10, and 15 groups as shown in the tables below. For each of these tests, the p-value is higher than $\alpha = 0.05$ (the lowest p-value is 0.4498) and quite stable, suggesting that the model provides a reasonable fit to the income non-response data.

Table 11: Hosmer and Lemeshow goodness of fit (GOF) test: `wtp.complete$INCOME_NONRESPONSE, wtp.model.int$fitted.values`

Test statistic	df	P value
2.56	3	0.4645

Table 12: Hosmer and Lemeshow goodness of fit (GOF) test: `wtp.complete$INCOME_NONRESPONSE`, `wtp.model.int$fitted.values`

Test statistic	df	P value
7.529	8	0.4808

Table 13: Hosmer and Lemeshow goodness of fit (GOF) test: `wtp.complete$INCOME_NONRESPONSE`, `wtp.model.int$fitted.values`

Test statistic	df	P value
12.97	13	0.4498

3. **Statistical learning: (12 marks)**

   Now we perform an exploratory analysis to try to identify the best set of predictors in predicting whether a respondent will not report a numeric value for income. Consider as predictors all variables in `wtp.complete` (other than the outcome of interest, `INCOME_NONRESPONSE`).

   a. **(4 marks) Find the optimal models identified by forward and backward selection algorithms. Report the predictors included in these optimal models. If these models are different, highlight how they differ, and explain why forward and backward selection algorithms may not arrive at the same optimal model.**

   With forward selection, we start with an empty model and sequentially add predictors that lead to the most significant improvement in fit until no predictors can be added that lead to a better model fit. The steps that were taken by the forward selection algorithm are as shown in the table below, leading to `PAY_WATER`, `ELECTRIC`, and `HEAD` being included in the final model.

```
# Load the "MASS" package to make use of the stepAIC() function.
library(MASS)
# Perform forward selection for models based on the specified predictors.
# Start with an empty model.
forward.selection.wtp <- stepAIC(glm(INCOME_NONRESPONSE ~ 1, family = "binomial", data =
wtp.complete), scope = list(upper = as.formula(paste("~",
paste(names(wtp.complete)[names(wtp.complete) != "INCOME_NONRESPONSE"], collapse = " +
"))), lower = ~1), direction = "forward", trace = FALSE)
```

Table 14: Steps taken by forward selection in adding predictors to the model.

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 selection starts with a full model and sequentially removes predictors that produce likelihood ratio tests with the highest (non-statistically significant) p-values until no more predictors can be removed. The steps that were taken by the backward selection algorithm are as shown in the table below, leading to the same model as produced using forward selection.

```
# Perform backward selection for models based on the specified predictors.
# Start with a full model.
backward.selection.wtp <- stepAIC(glm(as.formula(paste("INCOME_NONRESPONSE ~",
paste(names(wtp.complete)[names(wtp.complete) != "INCOME_NONRESPONSE"], collapse = " +
"))), family = "binomial", data = wtp.complete), scope = list(upper =
as.formula(paste("~", paste(names(wtp.complete)[names(wtp.complete) !=
"INCOME_NONRESPONSE"], collapse = " + "))), lower = ~1), direction = "backward", trace =
FALSE)
# Output the steps that were taken in the backward selection algorithm to produce the
final model.
pander(backward.selection.wtp$anova, caption = "Steps taken by backward selection in
removing predictors from the model.")
```

Table 15: Steps taken by backward selection in removing predictors from the model.

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 table below provides a comparison of the predictors that are included in the "best" models selected by forward selection and backward selection algorithms. We know that it is possible for forward and backward selection algorithms to arrive at different final models, as stepwise selection algorithms are "greedy" algorithms which simply take the optimal choice at each step, rather than exploring all possible subsets of predictors. In this case, however, forward and backward selection algorithms select the same set of predictors.

Table 16: Predictors included (✓) in optimal models selected by forward and backward selection algorithms.

Variable	Forward selection	Backward selection
TOWN		
SEX		
AGE		
EDUC		
HEAD	✓	✓
PAY_WATER	✓	✓
ELECTRIC	✓	✓
TIME_LENGTH		

b. **(4 marks) Find the optimal models identified by best subset selection using AIC and BIC as selection criteria. Report the predictors included in these optimal models. If these models are different, highlight how they differ, and explain why the criteria of AIC and BIC may lead to different "best" models.**

For best subset selection according to minimising AIC or minimising BIC, we must first construct a data frame that includes all of the predictors and has the response as the last variable. Our reduced dataset already has this structure, but the name of the last variable must be changed from `INCOME_NONRESPONSE` to `y`. The code below does this.

```r
# Construct a data frame to be used by the bestglm() function.
# The structure of this data frame is rigid with predictors first and the response
# being placed in the last column.
# Note that the response variable MUST be named 'y' in the data frame.
predictors.for.bestglm <- data.frame(wtp.complete)
names(predictors.for.bestglm)[ncol(predictors.for.bestglm)] <- "y"
```

The tables below show predictors included in the top five models according to the criteria of minimising AIC and minimising BIC.

```r
# Load the "bestglm" library to make use of the bestglm() function.
library(bestglm)
# Find the best logistic regression model based on the predictors according to the
# criterion of minimising AIC.
best.logistic.AIC <- bestglm(Xy = predictors.for.bestglm, family = binomial, IC = "AIC",
method = "exhaustive")
# Show the top five models in terms of minimising AIC.
panderOptions("table.continues", "")
pander(best.logistic.AIC$BestModels, caption = "Variables selected in the top five models
according to minimising AIC.")
```

Table 17: Variables selected in the top five models according to minimising AIC. (continued below)

TOWN	SEX	AGE	EDUC	HEAD	PAY_WATER	ELECTRIC	TIME_LENGTH	Criterion
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	1328
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	1329
FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	1329
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	1330
FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	1330

```r
# Find the best logistic regression model based on the predictors according to the
# criterion of minimising BIC.
best.logistic.BIC <- bestglm(Xy = predictors.for.bestglm, family = binomial, IC = "BIC",
method = "exhaustive")
# Show the top five models in terms of minimising BIC.
pander(best.logistic.BIC$BestModels, caption = "Variables selected in the top five models
according to minimising BIC")
```

Table 19: Variables selected in the top five models according to minimising BIC (continued 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 table below provides a comparison of the “best” subset of predictors produced by an exhaustive subset search and using the criteria of minimising AIC and minimising BIC. For reference, the predictors included in optimal models using forward and backward selection algorithms are also shown. The optimal model based on minimising AIC matches the model that was selected using both forward and backward selection (i.e., all nine predictors), whereas the optimal model based on minimising BIC is the same except that it does not include `HEAD` (the last variable added using forward selection), an indicator for whether the respondent is the head of the household. We know that the only difference between AIC and BIC is in terms of the penalty term with the penalty term for BIC being increasingly larger than that of AIC as the sample size increases. (Here, we have a large sample size, so the penalty for BIC is much larger than that for AIC.) This means that the optimal set of predictors according to the criterion of minimising BIC will be a subset of the predictors chosen according to the criterion of minimising AIC. We note that, although the optimal model identified by minimising AIC matches the two models selected using stepwise selection algorithms, this need not be the case, as best subset selection does an exhaustive search of all possible combinations of predictors, whereas forward and backward selection do not.

Table 21: Predictors included (✓) in optimal models selected by forward and backward selection algorithms and best subset selection according to the criteria of minimising AIC and minimising BIC.

Variable	Forward selection	Backward selection	AIC	BIC
`TOWN`				
`SEX`				
`AGE`				
`EDUC`				
`HEAD`	✓	✓	✓	
`PAY_WATER`	✓	✓	✓	✓
`ELECTRIC`	✓	✓	✓	✓
`TIME_LENGTH`				

c. **(4 marks) Consider all possible combinations of the eight predictor variables for a cross-validation routine to select the optimal model(s) based on maximising area**

under the receiver operating characteristic curve (AUC). Use 50 repetitions of 10-fold cross-validation. If this model (or these models) differ from those identified as "best" in parts (a) and (b), explain why this may be the case.

In total, we fit $2^8 - 1 = 255$ possible models.

```
# Specify the indices of the variables to be considered in predictive models for income
non-response.
variable.indices <- 1 : (ncol(wtp.complete) - 1)

# Produce a matrix that represents all possible combinations of variables.
# Remove the first row, which is the null model (i.e., no predictors).
all.comb <- expand.grid(as.data.frame(matrix(rep(0 : 1, length(variable.indices)), nrow =
2)))[-1, ]
```

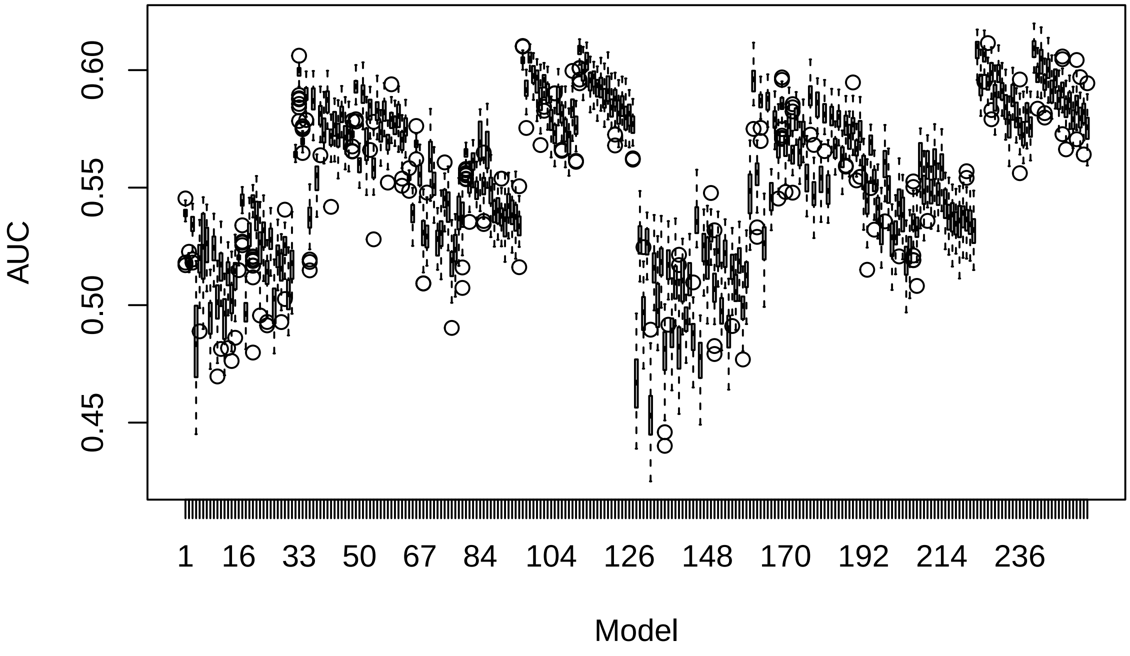
Code for calculating AUC for 50 repetitions of 10-fold cross-validation for these models is as shown below. Boxplots are produced to visualise the AUC for these repetitions of 10-fold cross-validation across the models, and the predictors contained in models that are within one standard error of the "best" model in terms of maximising AUC are also presented. Of the models within one standard error of the maximum mean AUC, the one that will have the maximum mean AUC can quite easily change if the cross-validation routine is re-run, although the optimal models will generally be the same. (This is because the `foreach` function does not have a simple mechanism to allow a seed to be set across multiple processors.) Consequently, the optimal model reported in the text may not necessarily match the R code output below (nor will it necessarily match your output either). For the criterion of maximising AUC, I have reported the models that performed best across three separate runs of the cross-validation routine.

```
#####################################
## Define functions for the cost   ##
## corresponding to the area under ##
## the ROC curve.                  ##
#####################################

area.under.curve <- function(r, p = 0)
{
            require(ROCR)

            pred <- prediction(p, r)
            auc <- performance(pred, measure = "auc")
            auc@y.values[[1]]
}
```

```
# Load the "doParallel" package to allow for parallel processing.
library(doParallel)
# Load the "foreach" package to allow for splitting loops.
library(foreach)
# Load the "boot" package to make use of the cv.glm() function .
library(boot)
# Set random number generator seed for replicability of results.
set.seed(1)

# Specify the number of repetitions of ten-fold cross-validation to carry out.
nrep <- 50

# Fire up 75% of cores for parallel processing.
nclust <- makeCluster(detectCores() * 0.75)
```

```
registerDoParallel(nclust)

################################
## Area under the ROC curve ##
################################

AUC.parallel <- foreach(i = 1 : nrep, .combine = "rbind", .packages = "boot") %:%
foreach(j = 1 : nrow(all.comb), .combine = "c") %dopar%
{
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])
}

# Shut down cores.
stopCluster(nclust)
```

```
################################
## Area under the ROC curve ##
################################

# View AUC according to model.
boxplot(AUC.parallel ~ matrix(rep(1 : nrow(all.comb), each = nrep), nrow = nrep), xlab =
"Model", ylab = "AUC")
```



```
# View all models within one SE of the best model.
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]]) #
Variable names
print(apply(AUC.parallel, 2, mean)[best.models.AUC[i]]) # AUC
cat("\n")
}
```

```
Model  1 :
[1] "PAY_WATER" "ELECTRIC"
[1] 0.6043317

Model  2 :
[1] "SEX"       "PAY_WATER" "ELECTRIC"
[1] 0.6052791

Model  3 :
[1] "HEAD"      "PAY_WATER" "ELECTRIC"
[1] 0.6077764

Model  4 :
[1] "PAY_WATER"   "ELECTRIC"    "TIME_LENGTH"
[1] 0.6082779

Model  5 :
[1] "SEX"         "PAY_WATER"   "ELECTRIC"    "TIME_LENGTH"
[1] 0.6063767

Model  6 :
[1] "HEAD"        "PAY_WATER"   "ELECTRIC"    "TIME_LENGTH"
[1] 0.6091461
```

The optimal models according to maximising AUC is similar to that identified by stepwise selection algorithms and best subset selection using AIC or BIC in that they include `PAY_WATER` and `ELECTRIC` (black ✓ in the table below) and one other predictor (either `HEAD`, `SEX`, or `TIME_LENGTH`, which are denoted by a red ✓ in the table below). It is important to recall that, for these other methods, there is no distinction between the training and test sets (i.e., the same data used to fit the model is used to assess model performance), and, even though there is a penalty for additional model complexity in the case of AIC and BIC, this does not guarantee that the approach is free of the problem of overfitting. Consequently, the optimal model based on maximising AUC need not match these other approaches.

Variable	Forward selection	Backward selection	AIC	BIC	AUC
`TOWN`					
`SEX`					✓
`AGE`					
`EDUC`					
`HEAD`	✓	✓	✓		✓
`PAY_WATER`	✓	✓	✓	✓	✓
`ELECTRIC`	✓	✓	✓	✓	✓
`TIME_LENGTH`					✓

Assignment total: 38 marks

References

Argent, J. 2009. “Household Income: Report on NIDS Wave 1.” 3. National Income Dynamics Study, University of Cape Town, Cape Town.

Biewen, M. 2001. “Item Non-Response and Inequality Measurement: Evidence from the German Earnings Distribution.” *Allgemeines Statistisches Archiv* 85: 409–25.

Dixon, J. 2005. *Comparison of Item and Unit Nonresponse in Household Surveys.* Bureau of Labor Statistics, Washington, D.C.

Fonseca, C. 2014. “The Death of the Communal Handpump? Rural Water and Sanitation Household Costs in Lower-Income Countries.” PhD thesis, Applied Sciences, Water Sciences, Cranfield University.

Juster, F. T., H. Cao, M. Perry, and M. Couper. 2006. “The Effect of Unfolding Brackets on the Quality of Wealth Data in HRS.” WP 2006-113. Michigan Retirement Research Center, University of Michigan, Ann Arbor.

Lillard, L., J. P. Smith, and F. Welch. 1986. “What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation.” *Journal of Political Economy* 94 (3): 489–506.

Moore, J. C., L. L. Stinson, and E. J. Welniak. 2000. “Income Measurement Error in Surveys: A Review.” *Journal of Official Statistics* 16 (4): 331–62.

Schräpler, J. P. 2004. “Respondent Behavior in Panel Studies: A Case Study for Income Nonresponse by Means of the German Socio-Economic Panel (SOEP).” *Sociological Methods & Research* 33: 118–56.