DATA303-DATA473 Assignment 4

Due: 11:59 PM Thursday 2 June 2022

Instructions

- Prepare your assignment using Rmarkdown.
- Submit your solutions in two files: an Rmarkdown file named `assignment4.Rmd` and the PDF named `assignment4.pdf` that results from knitting the Rmarkdown file.
- The YAML header of your Rmarkdown file must contain your name and ID number in the author field, and should have the output format set to `pdf_document`. For example:

```
---
title: "DATA303-DATA473 Assignment 4"
author: "Ryan Admiraal, 12345678"
date: "28 May 2021"
output: pdf_document
---
```

- While you are developing your code you may find it easiest to have the output set to `html_document` but change it to `pdf_document` when you submit.
- In your submission, embed any executable `R` code in code chunks, and make sure both the `R` code and the output is displayed correctly when you knit the document.
- If there are any `R` code errors, then the Rmarkdown file will not knit, and no output will be created at all. If you cannot get your code to work but want to show your attempted code, then put `error = TRUE` in the header of the `R` code chunk that is failing.

```
```{r, error = TRUE}
your imperfect R code
```
```

- Where appropriate, make sure you include code comments in the output within the Rmarkdown document to explain what lines or blocks of code are doing.
- **You will receive an email confirming your submission. Check the email to be sure it shows that both the Rmarkdown file and the PDF have been submitted.**

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 changed to point to the location of the file on your computer.

```
# 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.

   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.

   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?

   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.)**

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.

   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.

   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".

   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.

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.

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.

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.

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.

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.

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.