Tutorial letter 013/0/2024

Generalized Linear Models

STA4813

Year module

Department of Statistics

ASSIGNMENT 03 QUESTIONS

ASSIGNMENT 03
Unique Nr.: 859068
Fixed closing date: 05 July 2024

Instructions

(1) Use R Markdown to compile your solutions.

(2) Your solutions must have the full R codes and outputs of all questions.

(3) Discuss the relevant R outputs that are related to a specific question.

QUESTION [55]

The data `LabourForce.csv`, available in the "Additional Resources" folders on the module website, were extracted from the Quarterly Labour Force Survey of Quarter 1, 2021 conducted by the Statistics South Africa (Stats SA). Information on the survey and its metadata is available in the Stats SA Library Cataloguing-in-Publication (CIP) [1].

The focus of this assessment is on the variables presented in Table 1 below:

[1] Data Quarterly Labour Force Survey Quarter 1, 2021: Metadata/Statistics South Africa. Pretoria: Statistics South Africa, 2020 138pp. [Statistical Release: P0211 (2020)] ISBN: 978-0-621-49494-5, Quarterly

Table 1: Variables and their description in *LabourForce.csv.*

Question number	Variable name	Description
UQNO	Unique number	Unique household identifier allocated to each household
Q13GENDER	Gender	Gender of a household member, 1 = Male, 2 = Female
Q15POPULATION	Population_group	Population group of a person, 1 = African/Black, 2 = Coloured, 3 = Indian/Asian, 4 = White
Q16MARITALSTATUS	Marital_status	Marital status of a household member, 1 = Married, 2 = Living together like husband and wife, 3 = Widow/Widow 4 = Divorced or separated, 5 = Never married
Education_Status	Education Status	Derived variable [2] 1 = No schooling, 2 = Less than primary completed, 3 = Primary completed, 4 = Secondary not completed, 5 = Secondary completed, 6 = Tertiary, 7 = Other
Q24APDWRK	Paid_work	A household member aged 15 years and older who in the last week did any work for a wage, salary, commission or any payment in-kind, 1 = Yes, 2= No
Q14AGE	Age	Age of a household member[3]
Province	Province [4]	1 = Western Cape, 2 = Eastern Cape, 3 = Northern Cape, 4 = Free State, 5 = KwaZulu-Natal, 6 = North West, 7 = Gauteng, 8 =Mpumalanga, 9 = Limpopo
Geo_type_code	Geography type	1 = Urban, 2 = Traditional, 3 = Farms

[2]Derived from Question 1.7: What is the highest level of education that . . . has successfully completed?
[3] This is a derived variable indicating the age of the household member. Age of the household member was derived from question 1.4 of the questionnare. This question was asked for each member of the household. The instruction was to write the age in completed years to the nearest whole numbers and not in words. Thus, if a person was two years and six months, the instruction was to write the two completed years. For children aged less than a year, the instruction was to write 000.
[4] South African provinces as at December 2005 released by the Municipal Demarcation Board in January 2006.

The purpose of this assessment is to determine whether a household member aged 15 years and older who in the last week did any work for a wage, salary, commission or any payment, *Paid_work*, could be affected by the other seven variables (excluding variable UQNO).

(a) Plot the relationship between the response *Paid_work* and a household member aged 15 years and older (Hint: you should extract the data from Age variable). Explain why the plot does not work well. (3)

(b) Fit a binary regression with *Paid_work* as the response and the other seven variables as predictors, excluding variable UQNO. Report the residual deviance and associated degrees of freedom. Can this information be used to determine if this model fits the data? Explain. (10)

Using the model fitted in part (b):

(c) Use AIC as the criterion to determine the best subset of variables. (Use the `step` function.) (5)

(d) Plot the deviance residuals against the fitted values. What can be concluded from this plot? (5)

(e) Produce a binned residual plot as described in the Learning Unit 3 lecture notes. You will need to select an appropriate amount of binning. Comment on the plot. (3)

(f) Plot the binned residuals against *Age*, i.e.,the predictor. Comment on the plot. (3)

(g) Produce a normal QQ plot for the residuals. Interpret the plot. (3)

(h) Make a plot of the leverages. Interpret the plot. (3)

(i) Check the goodness of fit for model in part (b) by computing the Hosmer-Lemeshow statistic and associated p-value. What do you conclude? (5)

(j) Interpret the estimated coefficients for the fitted model. (15)

[55]