BTRY 6020 Homework V

NAME: student name

NETID: student NetID

DUE DATE: 8:40 am Friday March 31

Question 1.

Health officials wonder why some people get the flu shot while others don't. In a study designed to shed some light on this, researchers asked a random sample of patients if they had gotten aflu shot, recorded their age and gender, and also gave each a written questionnaire designed to evaluate their health awareness index. Data appear in Hwk5Q1DatSp17. Note here that Y = 1 means they received the flu shot and that males were coded as $X_3 = 1$, females coded as $X_3 = 0$.

A) Obtain the maximum likelihood estimators of β_o, β_1, β_2, and β_3. State the fitted regression function.

B) What is the estimated probability of getting the flu shot that a male clients aged 55 years with a health awareness score of 60?

C) Obtain the VIFs for the regression predictors. What conclusions can you reach from these statistics?

D) Get the standardized deviance residuals and plot against observation number. Does there appear to be any outliers?

E) Get the Cook's distance numbers and plot against observation number. Do there appear to be any influential outliers? If so, check their effects.

F) Can we drop Age and Gender if we keep the health awareness index in the model? State hypotheses, test statistic, p-value, and conclusions.

G) Install the package "bestglm". Visit the following website:

https://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf

to learn how to use this package. Don't forget the "library(bestglm)" command before you use it.

`i) Find the best model for getting a flu shot according to the BIC criteria`

`ii) Find the best models for a 0, 1, 2, and 3 predictors using the Subsets command`

`iii) Find the best model for getting a flu shot according to the AIC criteria`

`iv) Find the best models for a 0, 1, 2, and 3 predictors using the Subsets command`

v) What model from the above models evaluated would you choose for this situation? Explain BRIEFLY; you may include data from all parts of Question 1.

Question 2.

A disease outbreak has occurred in a certain city. Data have been collected on a random telephone survey of 196 people within city limits and the following data recorded: 1) Whether or not they have contracted the disease (Dis, =1 if they have, =0 if not), Age, Socioeconomic Status (SES, = 1 if upper, = 2 if middle, = 3 if lower), Sector of the city they live (Sect, either sector 1 or sector 2), and saving account status (Sav, = 1 if they have a savings account, = 0 if not). data appear in Hwk5Q2DatSp17

Part A) Develop a logistic regression model for predicting the probability of contracting this disease, using the above variables. Be sure to check for polynomial effects of significant quantitative variables as well as interactions between significant predictor variables. When finished, explicitly state your prediction equation. Be sure to show significant steps in model development, using simultaneous tests when you want to omit/test more than one predictor.

Part B) Give a 90% confidence interval for the probability that a 64 year old patient, with middle socioeconomic status and a savings account that lives in sector 2 of the city, contracts the disease.

Question 3.

Multiple cohorts of subjects, some non-smokers and others smokers, were observed for several years. The number of cases (NumCases) of lung cancer diagnosed in the different cohorts was recorded, in addition to the following predictor variables:

CigsperDay = Number of cigarettes smoked per day per individual in the cohort; Years = The number of years the individuals in the cohort had smoked.

Additionally, the total number of years in which individuals in each category were observed (summed over all individuals) was recorded in the column PersonYears. (For example, if a cohort had 50 people that had been observed for 20 years, that would be 50 x 20 = 1000 PersonYears.) Data appear in Hwk5Q3DatSp17.

A) Write down a Poisson regression model where the mean number of cases of observed lung cancer cases per cohort are a function of CigsperDay and Years. Your model should start like "μ = .", NOT "log(μ) = .".

B) Fit the model above; include summary output. State your model of the estimated mean with the maximum likelihood estimators included.

C) Do a deviance goodness-of-fit test on your model; state hypotheses, test statistic, p-value, and conclusions.

D) Does it make sense for your mean in Part A above to be proportional to the variable PersonYears? Explain briefly.

E) Write down a Poisson regression model where the mean number of cases of observed lung cancer cases per cohort are a function of CigsperDay and Years, but are also proportional to PersonYears. Your model should start like "μ = .", NOT "log(μ) = ."

F) Fit the above model; include summary output. Perform a deviance goodness-of-fit test on this model; state hypotheses, test statistic, p-value, and conclusions.