STAT 561 - Homework 2

Instructions

1. *Due Monday, April 14th at the 11:59pm. Any submission after that (24hrs after) will be graded out of 50%.*

2. *Your submission should include a* ***pdf*** *from your generated R markdown, and your R markdown file.*

3. *Make sure to highlight the part of the output that have the information you will be providing as answers.*

4. *This work should be done as a group. Only one person in the group should submit the work with the names of all the people in the group on the document*

---

1. **Absenteesim data** In this work, we tackle the challenge of absenteeism in today's dynamic work environment using logistic regression. Leveraging a detailed dataset which was created with records of absenteeism at work from July 2007 to July 2010 at a courier company in Brazil, we aim to predict and understand the patterns of absenteeism. The features to be used are:

Month of absence, Day of the week, Seasons, Transportation expense, Distance from Residence to Work, Service time, Age, Work load Average/day, Education, Son, Pet, Weight, Height, Body mass index. Be sure to make the categorical variables factors in R.

Your task is to develop a multinomial logistic regression model to predict the level of absenteeism (Low, Moderate, High) of an employee based on various predictors. The absenteeism time in hours should be categorized into three groups: Low: 0 – 20 hours, Moderate: 21 – 40 hours, and High Absenteeism: $>$ 40 hours. You can utilize the following code:

```
library(dplyr)
library(readxl)

absent= read_excel("Absenteeism_at_work.xls")
names(absent) #column names
View(absent)

#recode absenteeism

absent <- absent %>%
mutate(absenteeism = case_when(
Absenteeism_time_in_hours >= 0 & Absenteeism_time_in_hours <= 20 ~ "Low",
Absenteeism_time_in_hours > 20 & Absenteeism_time_in_hours <= 40 ~ "Moderate",
Absenteeism_time_in_hours > 40 ~ "High"
))
View(absent)
```

Exploratory data analysis

Use R to create any of the following to help answer the following questions: histograms, box plots, scatter plots, correlations and or correlation matrices, bar graphs, summary(), etc.

(a) How is the 'Absenteeism time in hours' distributed? Are there any noticeable patterns or outliers?

(b) What is the distribution of ages among the employees? Are certain age groups more prevalent?

(c) Is there a correlation between the distance from residence to work and absenteeism time?

(d) How does the work load average per day relate to absenteeism? Are higher workloads associated with more or less absenteeism?

(e) Analyze the absenteeism based on education levels. Do certain education levels correlate with higher or lower absenteeism?

(f) Which variables show the strongest correlation with absenteeism time in hours? How might these influence your logistic regression model?

(g) Are there any unexpected correlations or findings that challenge common assumptions about workplace absenteeism?

(h) Does service time (duration of service in the company) have any impact on the absenteeism rate?

(i) Examine if day of the week has any influence on absenteeism – are certain days more prone to absences?

(j) Identify any outliers in the data set. What could be the reasons for these anomalies, and how might they affect the analysis?

Note: Generally, an approach in choosing the features to use in a model is to use your EDA results, based on each variable's relationship with the response

Logistic Regression Analysis Instructions

(a) Build a logistic regression model using the recoded absenteeism categories (Low, Moderate, High) as the response variable. Ensure that categorical variables are appropriately handled.

(b) Interpret the coefficients of the variables son and weight.

(c) Use backward selection to decide which predictor variables enter should be kept in the regression model.

(d) Interpret some (at least 2) of your model's findings in a practical workplace context. Formulate recommendations for an organization based on these findings.

Variable description

1. Month of absence
2. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
3. Seasons (summer (1), autumn (2), winter (3), spring (4))
4. Transportation expense
5. Distance from Residence to Work (kilometers)
6. Service time
7. Age
8. Work load Average/day
9. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
10. Son (number of children)
11. Pet (number of pet)
12. Weight
13. Height
14. Body mass index
15. Absenteeism time in hours (target)

2. **Flu shots data** A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications to get a flu shot in time for protection against an expected flu epidemic. In a pilot follow-up study, 159 clients were randomly selected and asked whether they actually received a flu shot. A client who received a flu shot was coded $Y = 1$, and a client who did not receive a flu shot was coded $Y = 0$. In addition, data were collected on their age (X_1) and their health awareness. The latter data were combined into a health awareness index (X_2), for which higher values indicate greater awareness. Also included in the data was client sex, where males were coded $X_3 = 1$ and females were coded $X_3 = 0$.

(a) Create a scatterplot matrix of the data. What are your observations?

(b) Fit a multiple logistic regression to the data with the three predictors in first order terms.

(c) State the fitted regression equation.

(d) Obtain $\exp(\hat{\beta}_1)$, $\exp(\hat{\beta}_2)$, $\exp(\hat{\beta}_3)$ and interpret these numbers.

(e) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot?

(f) Using the Wald test, determine whether X_3 , client gender, can be dropped from the regression model; use $\alpha = 0.05$.

(g) Use forward selection to decide which predictor variables enter should be kept in the regression model.

(h) Use backward selection to decide which predictor variables enter should be kept in the regression model. How does this compare to your results in part (f)?

(i) How would you interpret $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_3$