**Quiz 7 – Logistic Regression**

Content: glm(); predict()

Data files <churn.csv> is provided on Canvas.

Please open the R code file attached <Quiz7.R> and run the code to answer the questions below.

All the log calculation mentioned in this quiz is the natural log (base e).

Note: to run the code file smoothly, you might need to:

Download the data file above.

Set working directory in R to the folder that contains the data file: setwd()

**Q1.** Model M1 generated in the R code file is a

Linear Regression model

Logistic Regression model  
Naïve Bayes model

KNN model

Ans: Linear Regression model

Explanation: Model M1 was generated by function glm(), however, family = binomial was not specified. Hence, by default settings of glm(), it will form a linear model.

**Q2.** How many coefficients does model M2 have (including the intercept)?

3, 4,5,6

Answer: 5

Explanation: Model M2 was generated with Churned be the response, and it depends on ALL other columns (4 columns) in the data frame “data”. Each column will need one coefficient. Hence, 4 coefficients for 4 variables with 1 intercept 🡪 5 in total.

Note: Married is categorical with 2 categories, hence, it needs one coefficient only, not two.

**Q3.** Variable “Married” is NOT very significant in model M2. True or False?

Ans: True. Its p-value is 0.33, quite large. It indicates the weak contribution of this variable in the model.

**Q4.** Consider model M3. From the output of summary(M3), if we let p = prob(Churned = 1), and p-hat be the estimation of p, then we could write the fitted equation of model M3 as

1. Log(p-hat) = 3.503 – 0.157\*Age + 0.382\*Churned\_contacts
2. Log [ p-hat/(1-p-hat)] = 3.503 – 0.157\*Age + 0.382\*Churned\_contacts
3. p-hat = 3.503 – 0.157\*Age + 0.382\*Churned\_contacts
4. p = 3.503 – 0.157\*Age + 0.382\*Churned\_contacts
5. p-hat / (1 – p-hat) = 3.503 – 0.157\*Age + 0.382\*Churned\_contacts

Ans: Option B.

**Q5.** Consider using model M3 to predict for a customer at age 50 and have 5 churned contacts. What is the output of the R code in line 22?

predict(M3, newdata = data.frame(Age = 50, Churned\_contacts = 5), type = 'response')

1. The value of Churned for that customer.
2. The predicted odds that the customer will churn.
3. The predicted probability that the customer will churn.
4. The predicted ratio between probability of churn and probability of not churn for that customer.

Ans: Option C.

Explanation: type = ‘response’ is specified which requires R to return the probability of response = 1. That is the probability predicted by model M3.

**Q6.** This question has one or more than one answer that is correct.

The coefficient -0.157 (3dp) in the summary(M3) output for variable Age means

1. As Age increases by 1 unit (1 year), the probability of customer to churn will decrease by 0.157.
2. As Age increase by 1 unit (1 year), the log odds of customer to churn will decrease by 0.157.
3. As Age increase by 1 unit (1 year), the odds of customers to churn will decrease by 0.157.
4. As Age increase by 1 unit (1 year), the odds of customer to churn will decrease by 1.17 times.

Ans: both options B and D are correct.

Explanation: Refer to the correct answer of Q4 to get why B is correct.

To get an idea on why D is correct, we consider equation below.

log(y) = 5 \* Age

When Age increase by 1 then log(y) will increase by 5, hence the value of y will increase by an exponential of 5, exp(5).

**Q7.** When we want to fit a logistic model, the response variable could be categorical with "0" and "1" or could be numeric with 0 and 1. True or False?

Ans: True.

From the R code file, you can see that "Churned" is numeric with 0 and 1.

glm() can fit a logistic model for this response.

If you change response "Churned" into categorical before fitting logistic model, it still works (you can try with the code below):

data$Churned = as.factor(data$Churned)  # change "Churned" into categorical, and fit model M2

M2<- glm(Churned ~ . , data = data, family = binomial)   # no error  
summary(M2)  # it still works as usual.

An extra note: if the response variable is categorical with characters "0" and "1" then glm() can fit a logistic model for us. HOWEVER, if the response has characters like "yes" and "no", then glm() cannot fit a logistic model.