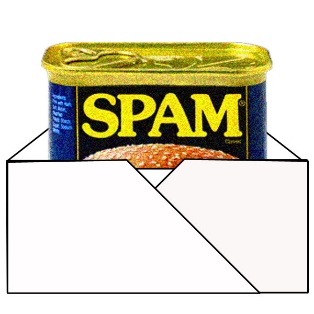
**Module Assignment**

**Module 8**

**QMB-6304 Analytical Methods for Business**



Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load the file “Module 8 Assignment Data Set.xlsx” into R. This file contains information on 3914 emails and whether those emails were judged to be spam. This will be you master data set. Variables included are:
   1. ROW: A unique identifier for each row of the dataframe.
   2. SPAM: A binary variable identifying whether the message was spam. (1 = yes)
   3. TO\_MULTIPLE: A variable indicating whether more than one recipient was listed in the “To” field of the email. (1 = yes)
   4. IMAGE: A numerical variable showing how many images were in the email.
   5. DOLLAR: A numerical variable showing how many times “$” or the word “dollar” appeared in the email.
   6. WINNER: A variable indicating whether the word “winner” appeared at least once in the email. (1 = yes)
   7. INHERIT: A numerical variable showing how many times the word “inherit” appears in the email.
   8. PASSWORD: A numerical variable showing how many times the word “password” appears in the email.
   9. FORMAT: An variable indicating whether the email contains special formatting such as bolding, tables, or links. (1 = yes)
   10. RE\_SUBJ: A variable indicating whether “Re:” was included in the subject line. (1 = yes)
   11. URGENT\_SUBJ: A variable indicating whether “Urgent” was included in the subject line. (1 = yes)
   12. EXCLAIM\_SUBJ: A variable indicating whether “!” was included in the subject line. (1 = yes)
2. Using the numerical portion of your U number as a random number seed, take a random sample of 750 cases from the full data set using the method presented in class. This will be your primary data set.

**Analysis (Using the Primary Data Set)**

1. Parameterize a full logistic regression model with SPAM as the dependent and all other variables as independent (excluding ROW).
2. Using the *summary()* command report the results of the final recommended model from Step 1.
3. State whether you believe the Residual Deviance of your model is markedly different from the Null Deviance.
4. Given your model from Part 1 and ignoring the p values, which variable will have the greatest influence in increasing the modeled probability that an email will be spam?
5. Given your model from Part 1 and ignoring the p values, which variable will have the greatest influence in decreasing the modeled probability that an email will be spam?
6. You likely found that R would not determine a beta coefficient in your model for the URGENT\_SUBJ variable. If this happened in your model, what caused this result?
7. Parameterize a new logistic regression model with the following variables as independents: TO\_MULTIPLE, DOLLAR, INHERIT, PASSWORD, and FORMAT.
8. Use the *expand.grid()* command develop a prediction file with all independent variables in the Step 7 model. For binary independent variables use the *unique()* qualifier. For numerical (continuous) independent variables use the *quantile()* qualifier and set test levels at the 25th, 50th, 75th, and 100th percentiles for the variables as appearing in your reduced data set. Examples of these qualifiers can be found in the video PowerPoints. Calculate and show independent variable values and predicted probabilities for ONLY the first five cases appearing in your prediction file.
9. Based on your predictions generated in Step 7, state the maximum and minimum predicted probabilities generated and the independent variable values which resulted in those predictions.

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above preprocessing and analysis instructions and 2) the results of those instructions and needed written interpretations. Results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**

**Preprocessing:**

**#Varun Teja Kolluru**

**#preprocessing**

**#1**

**rm(list=ls())**

**library(rio)**

**library(moments)**

**my\_data=import("6304 Module 8 Assignment Data Set.xlsx")**

**attach(my\_data)**

**#2**

**set.seed(97)**

**my\_sample = my\_data[sample(1:nrow(my\_data),750),]**

Using ‘rm’ command we are clearing all the variables and vectors in the environment window. All the required packages are imported into our R project. Data is imported using the import command.

As per the 2nd question in preprocessing, a random seed with my last 2 digits of U number is set and a sample of 750 observations are taken for the analysis.

**Analysis:**

1. Parameterize a full logistic regression model with SPAM as the dependent and all other variables as independent (excluding ROW).

**RCode:**

**#analysis**

**#1**

**output = glm(spam ~ to\_multiple+ image+ dollar+ winner+ inherit+**

**password +format+re\_subj+ urgent\_subj + exclaim\_subj,**

**family=binomial, data = my\_sample)**

‘glm’ is the command in R to get the logistic regression and SPAM is used as an dependent variable and all the other variables are used as independent variables expect ROW and we used ‘tilda’ operator in between.

1. Using the *summary()* command report the results of the final recommended model from Step 1.

**RCode:**

**#2**

**summary(output)**

**Output in Console Window:**

> #2

> summary(output)

Call:

glm(formula = spam ~ to\_multiple + image + dollar + winner +

inherit + password + format + re\_subj + urgent\_subj + exclaim\_subj,

family = binomial, data = my\_sample)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.32819 -0.40887 -0.28699 -0.00006 3.08200

Coefficients: (1 not defined because of singularities)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.89518 0.21407 -4.182 2.89e-05 \*\*\*

to\_multiple -2.81955 0.74679 -3.776 0.000160 \*\*\*

image -12.39872 1353.09419 -0.009 0.992689

dollar -0.03392 0.03720 -0.912 0.361738

winneryes 2.33669 0.67656 3.454 0.000553 \*\*\*

inherit 0.40038 0.27493 1.456 0.145306

password -1.20373 1.05432 -1.142 0.253572

format -1.54457 0.29642 -5.211 1.88e-07 \*\*\*

re\_subj -17.73905 728.36941 -0.024 0.980570

urgent\_subj NA NA NA NA

exclaim\_subj 0.51863 0.48279 1.074 0.282721

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 446.83 on 749 degrees of freedom

Residual deviance: 343.77 on 740 degrees of freedom

AIC: 363.77

Number of Fisher Scoring iterations: 18

Intercept, to\_multiple, winnervers, and format are very significant.

1. State whether you believe the Residual Deviance of your model is markedly different from the Null Deviance.

Yes, the residual Deviance of the model is markedly different from the Null Deviance.

The difference between Null Deviance and Residual Deviance or Chi Square value is 103.06.

The Chi square value is large and good. So that, the logistic model is good fit to the data.

Null deviance shows how well response variable (spam) is predicted by a model that includes only intercept and On the other hand, residual deviance shows how well response variable (spam) is predicted by a model that includes independent variables.

For better model performance expected when the Deviance values are lower and the difference between Null deviance and residual deviance should be large.

But theoretically, null deviance value is greater than that of residual deviance.

1. Given your model from Part 1 and ignoring the p values, which variable will have the greatest influence in increasing the modeled probability that an email will be spam?

From the Summary of the output of logistic regression, re\_subj has the highest beta coefficient which is 17.73.

This variable has the highest influence in increasing the model probability that email will be spamed and have negative sign for the re\_subj variable.

1. Given your model from Part 1 and ignoring the p values, which variable will have the greatest influence in decreasing the modeled probability that an email will be spam?

Dollar has the lowest beta coefficient which is -0.03392 and is negatively associated to the spam variable. It has least influence in predicting if email is spam.

1. You likely found that R would not determine a beta coefficient in your model for the URGENT\_SUBJ variable. If this happened in your model, what caused this result?

The coefficient of URGENT\_SUBJ in the output is shown as NA and this is because the Independent Variable URGENT\_SUBJ does not add any information to the model. URGENT\_SUBJ variable is linearly dependent on other independent variables in the model.

1. Parameterize a new logistic regression model with the following variables as independents: TO\_MULTIPLE, DOLLAR, INHERIT, PASSWORD, and FORMAT.

**RCode:**

**#7**

**new\_output=glm(spam~to\_multiple+dollar+inherit+password+format, data = my\_sample,family=binomial)**

**summary(new\_output)**

**coefficients(new\_output)**

**Output in Console Window:**

#7

> new\_output=glm(spam~to\_multiple+dollar+inherit+password+format, data = my\_sample,family=binomial)

> summary(new\_output)

Call:

glm(formula = spam ~ to\_multiple + dollar + inherit + password +

format, family = binomial, data = my\_sample)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.7153 -0.3621 -0.3621 -0.2151 3.2310

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.232501 0.197454 -6.242 4.32e-10 \*\*\*

to\_multiple -2.522204 0.736223 -3.426 0.000613 \*\*\*

dollar -0.007345 0.031567 -0.233 0.816010

inherit 0.459437 0.259683 1.769 0.076856 .

password -0.642171 0.724241 -0.887 0.375251

format -1.459459 0.277771 -5.254 1.49e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 446.83 on 749 degrees of freedom

Residual deviance: 401.23 on 744 degrees of freedom

AIC: 413.23

Number of Fisher Scoring iterations: 7

**Intercept, to\_multiple and format are very good significate to the model.**

**The difference between Null deviance and Residual Deviance is 45.6.**

**There is no change in Null Deviance value nut the residual deviance value is increased a lot and the chi square value has decreased significantly.**

> coefficients(new\_output)

(Intercept) to\_multiple dollar inherit password format

-1.232500894 -2.522203562 -0.007345087 0.459436671 -0.642170917 -1.459458707

**Below are the coefficients for the given variables and to\_multiple variable still has high significance.**

1. Use the *expand.grid()* command develop a prediction file with all independent variables in the Step 7 model. For binary independent variables use the *unique()* qualifier. For numerical (continuous) independent variables use the *quantile()* qualifier and set test levels at the 25th, 50th, 75th, and 100th percentiles for the variables as appearing in your reduced data set. Examples of these qualifiers can be found in the video PowerPoints. Calculate and show independent variable values and predicted probabilities for ONLY the first five cases appearing in your prediction file.

**RCode:**

**#8**

**pred = expand.grid(dollar=quantile(dollar,c(.25,.50,.75,1)),**

**password=quantile(password,c(.25,.50,.75,1)), to\_multiple=unique(my\_sample$to\_multiple),**

**inherit=quantile(inherit,c(.25,.50,.75,1)), format=unique(my\_sample$format))**

**reg\_nf = glm(spam ~ to\_multiple+dollar+inherit+password+format,data=my\_sample,family='binomial')**

**summary(reg\_nf)**

**coefficients(reg\_nf)**

**confint(reg\_nf)**

**prob\_predict=round(predict(reg\_nf,newdata=pred,type="response"),4)**

**spam\_predict=cbind(pred,prob\_predict)**

**spam\_predict[1:5,]**

**Output in Console Window:**

**#8**

**> pred = expand.grid(dollar=quantile(dollar,c(.25,.50,.75,1)),**

**+ password=quantile(password,c(.25,.50,.75,1)), to\_multiple=unique(my\_sample$to\_multiple),**

**+ inherit=quantile(inherit,c(.25,.50,.75,1)), format=unique(my\_sample$format))**

**> reg\_nf = glm(spam ~ to\_multiple+dollar+inherit+password+format,data=my\_sample,family='binomial')**

**> summary(reg\_nf)**

**Call:**

**glm(formula = spam ~ to\_multiple + dollar + inherit + password +**

**format, family = "binomial", data = my\_sample)**

**Deviance Residuals:**

**Min 1Q Median 3Q Max**

**-0.7153 -0.3621 -0.3621 -0.2151 3.2310**

**Coefficients:**

**Estimate Std. Error z value Pr(>|z|)**

**(Intercept) -1.232501 0.197454 -6.242 4.32e-10 \*\*\***

**to\_multiple -2.522204 0.736223 -3.426 0.000613 \*\*\***

**dollar -0.007345 0.031567 -0.233 0.816010**

**inherit 0.459437 0.259683 1.769 0.076856 .**

**password -0.642171 0.724241 -0.887 0.375251**

**format -1.459459 0.277771 -5.254 1.49e-07 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**(Dispersion parameter for binomial family taken to be 1)**

**Null deviance: 446.83 on 749 degrees of freedom**

**Residual deviance: 401.23 on 744 degrees of freedom**

**AIC: 413.23**

**Number of Fisher Scoring iterations: 7**

**> coefficients(reg\_nf)**

**(Intercept) to\_multiple dollar inherit password format**

**-1.232500894 -2.522203562 -0.007345087 0.459436671 -0.642170917 -1.459458707**

**> confint(reg\_nf)**

**Waiting for profiling to be done...**

**2.5 % 97.5 %**

**(Intercept) -1.63395923 -0.85743146**

**to\_multiple -4.35146471 -1.31180592**

**dollar -0.08193968 0.04473803**

**inherit 0.00342806 1.21498049**

**password -3.22900103 0.10459404**

**format -2.01048198 -0.91755041**

**Warning messages:**

**1: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**2: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**3: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**4: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**5: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**6: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**7: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**8: glm.fit: fitted probabilities numerically 0 or 1 occurred**

**> prob\_predict=round(predict(reg\_nf,newdata=pred,type="response"),4)**

**> spam\_predict=cbind(pred,prob\_predict)**

**> spam\_predict[1:5,]**

**dollar password to\_multiple inherit format prob\_predict**

**1 0 0 0 0 0 0.2257**

**2 0 0 0 0 0 0.2257**

**3 0 0 0 0 0 0.2257**

**4 64 0 0 0 0 0.1541**

**5 0 0 0 0 0 0.2257**

1. Based on your predictions generated in Step 7, state the maximum and minimum predicted probabilities generated and the independent variable values which resulted in those predictions.

**RCode:**

**#9**

**spam\_predict[which.max(spam\_predict$prob\_predict),]**

**spam\_predict[which.min(spam\_predict$prob\_predict),]'**

**Output in Console Window:**

**> #9**

**> spam\_predict[which.max(spam\_predict$prob\_predict),]**

**dollar password to\_multiple inherit format prob\_predict**

**97 0 0 0 9 0 0.948**

**> spam\_predict[which.min(spam\_predict$prob\_predict),]'**

**dollar password to\_multiple inherit format prob\_predictions**

**16 64 28 0 0 1 0**

When dollar, password, to\_multiple, format =0, inherit = 9; the max probability found is 0.948

When dollar = 64, password=28, to\_multiple=inherit=0, format=1, minimum probability is 0