**Module Assignment**

**Module 8**

**QMB-6304 Foundations of Business Statistics**

**#Debjani Sarma**

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

**> library(rio)**

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.xlsx” into R. This file contains information on 20,807 employees of the City of Chicago who received overtime pay during 2016. This will be your master data set. Variables included are:
   1. department.name: A character variable identifying the department in which the employee works. This data set includes only the five largest (by headcount) in the City of Chicago government in 2016.
   2. employee.name: The name of the employee.
   3. title: the job title of the employee
   4. january through december: the amount of overtime pay the employee received in each of the 12 months in 2016.
   5. total: the total amount of overtime pay the employee received in 2016.
   6. nummos: the number of months in 2016 in which the employee received overtime pay.
   7. over5000: a binary variable indicating whether the employee earned $5000 or more in overtime pay during 2016.

**> master.emp = import("6304 Module 8 Assignment Data.xlsx")**

1. Using the numerical portion of your U number as a random number seed, take a random sample of 4500 cases from the master data set using the method presented in class. This will be your primary data set for analysis.

**> set.seed(24173877)**

**> mydata = master.emp[sample(1:nrow(master.emp),4500),]**

**Analysis**

Using your primary data set:

1. Show the results of the str() command.

**> str(mydata)**

**'data.frame': 4500 obs. of 18 variables:**

**$ department.name: chr "Streets and Sanitation" "Water Management" "Police" "Police" ...**

**$ employee.name : chr "Williams, Darnell" "Jordan, Marvin R" "Collins, Brian" "Bacoulis, Daniel E" ...**

**$ title : chr "Pool Motor Truck Driver" "Motor Truck Driver" "Police Officer" "Police Officer" ...**

**$ january : num 3550 2088 NA 1627 NA ...**

**$ february : num 2864 2096 NA 1828 1506 ...**

**$ march : num 531 1631 NA 1553 NA ...**

**$ april : num 742 1510 NA 1553 NA ...**

**$ may : num 207 2105 NA 1553 NA ...**

**$ june : num 203 1078 412 1553 NA ...**

**$ july : num 70.1 380.9 375.8 1552.8 94.1 ...**

**$ august : num 26.3 183.9 NA 744.1 NA ...**

**$ september : num 210 718 NA NA NA ...**

**$ october : num 210 849 NA NA NA ...**

**$ november : num 140 631 NA NA NA ...**

**$ december : num 543 447 NA NA NA ...**

**$ total : num 9297 13718 787 11963 1600 ...**

**$ nummos : num 12 12 2 8 2 3 12 2 12 2 ...**

**$ over5000 : num 1 1 0 1 0 0 1 0 1 0 ...**

1. Parameterize a logistic regression model with over5000 as the dependent and department.name and nummos independent variables. Report the results of the model using the summary() command.

**> my.logreg = glm(over5000 ~ department.name + nummos, data = mydata, family = binomial)**

**> summary(my.logreg)**

**Call:**

**glm(formula = over5000 ~ department.name + nummos, family = binomial,**

**data = mydata)**

**Coefficients:**

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

**(Intercept) -4.04676 0.19532 -20.719 < 2e-16 \*\*\***

**department.nameFire 1.05781 0.18839 5.615 1.97e-08 \*\*\***

**department.namePolice -0.06233 0.17908 -0.348 0.7278**

**department.nameStreets and Sanitation -1.60539 0.21021 -7.637 2.22e-14 \*\*\***

**department.nameWater Management -0.54782 0.23180 -2.363 0.0181 \***

**nummos 0.68509 0.01932 35.464 < 2e-16 \*\*\***

**---**

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

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

**Null deviance: 6211.1 on 4499 degrees of freedom**

**Residual deviance: 3018.8 on 4494 degrees of freedom**

**AIC: 3030.8**

**Number of Fisher Scoring iterations: 6**

The intercept is the baseline of over5000 when all the independent variables are 0.

Having a positive coefficient (nummos and department.nameFire) indicates the probability of employee earing over $5000 increases, which all negative coefficients (department.namePolice, department.nameStreets and Sanitation, department.nameWater Management) indicates the probability of employee earning over $5000 decreases.

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

Based on the summary of logistic regression :

Null deviance : 6211.1 and Residual deviance : 3018.8

The Null Model represents model with just intercept (no independent variables)

Since the Residual Deviance is much lower than the Null deviance, it indicates that the model with independent variables (department.name and nummos) explains better variation in outcome (over5000) than the null model.

Yes, I believe the Residual Deviance of your model is markedly different from the Null Deviance.

1. Given your model from Part 2 and ignoring p values, which variable or variable/level will have the greatest influence in increasing the modeled probability that an employee earned $5000 or more in 2016?

Based on the summary of logistic regression , the variables department.nameFire and nummos have positive coefficients of 1.05781 and 0.68509 respectively. These variables influence in increasing the modeled probability that an employee earned $5000 or more in 2016.

Having nummos as positive coefficient suggests that each additional month employee received overtime, the probability of earning $5000 or more increased.

Having department.nameFire as positive coefficient suggests that employee in fire department had the probability of earning $5000 or more compared to other departments.

**department.nameFire** with the highest positive coefficient of 1.05781 will have the greatest influence in increasing the modeled probability that an employee earned $5000 or more in 2016.

1. Given your model from Part 2 and ignoring p values, which variable will have the greatest influence in decreasing the modeled probability that an employee earned $5000 or more in 2016?

Based on the summary of the logistic regression, the variables department.namePolice ,

department.nameStreets and Sanitation ,department.nameWater Management - all have negative coefficients, suggesting influence in decreasing the modeled probability that an employee earned $5000 or more in 2016. However, department.nameStreets and Sanitation has a negative coefficient of -1.60539, which is the most negative of all variables. Therefore, **department.nameStreets and Sanitation** will have the greatest influence in decreasing the modeled probability that an employee earned $5000 or more in 2016.

1. Using the *expand.grid()* command develop a prediction file with all independent variables in the Step 1 model. For independent variables in this case use the *unique()* qualifier. R will by default calculate predicted probabilities to many decimal places. For convenience in reporting round your stored predictions to only 4 decimal places. Show the predicted probabilities for ONLY the first ten cases appearing in your prediction file.

**> prediction = expand.grid(department.name = unique(mydata$department.name), nummos = unique(mydata$nummos))**

**> prediction$pred\_prob = round(predict(my.logreg, newdata = prediction, type = "response"),4)**

**> head(prediction, 10)**

**department.name nummos pred\_prob**

**1 Streets and Sanitation 12 0.9288**

**2 Water Management 12 0.9741**

**3 Police 12 0.9839**

**4 Fire 12 0.9947**

**5 Aviation 12 0.9848**

**6 Streets and Sanitation 2 0.0136**

**7 Water Management 2 0.0383**

**8 Police 2 0.0607**

**9 Fire 2 0.1654**

**10 Aviation 2 0.0644**

1. Based on your predictions generated in Step 6, find the maximum and minimum predicted probabilities generated. State the values of the independent variables for these max and min cases.

**> prediction.max = prediction[which.max(prediction$pred\_prob), ]**

**> prediction.max$department.name**

**[1] Fire**

**Levels: Streets and Sanitation Water Management Police Fire Aviation**

**> prediction.max$nummos**

**[1] 12**

**> prediction.min = prediction[which.min(prediction$pred\_prob), ]**

**> prediction.min$department.name**

**[1] Streets and Sanitation**

**Levels: Streets and Sanitation Water Management Police Fire Aviation**

**> prediction.min$nummos**

**[1] 0**

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is “Assignment 5” of our course and your name as it appears in Canvas. Your code chunks and analysis 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.**