# Dublin Business School

# Assessment Brief

# Assessment Details

|  |  |
| --- | --- |
| Module Title: | Statistics for Data Analytics |
| Module Code: | B9DA101 |
| Module Leader: | Dr Shahram Azizi |
| Stage (if relevant): |  |
| Assessment Title: | CA two |
| Assessment Number (if relevant): |  |
| Assessment Type: |  |
| Restrictions on Time/Length : | Submission before deadline |
| Individual/Group: |  |
| Assessment Weighting: |  |
| Issue Date: |  |
| Hand In Date: | Before final exam |
| Planned Feedback Date: |  |
| Mode of Submission: | Online |

**Guideline:**

* This CA assesses students on core concept in Hypotheses tests, GLM analytics, and Bayesian analytics.
* All questions are mandatory.
* Use R/Rstudio to solve questions and perform analytics.
* Any submission after deadline will not be considered and scored.

**Question 1**

Consider a relational dataset and specify your input and output variables, then:

* 1. Train the model using 80% of this dataset and suggest an appropriate GLM to model **ouput** to **input** variables.

**(10 Marks)**

* 1. Specify the significant variables on the **output** variable at the level of 𝛼=0.05 and explore the related hypotheses test. Estimate the parameters of your model.

**(10 Marks)**

* 1. Predict the output of the test dataset using the trained model. Provide the functional form of the optimal predictive model.

**(10 Marks)**

* 1. Provide the confusion matrix and obtain the probability of correctness of predictions.

**(5 Marks)**

**(Total: 35 Marks)**

***Question 1 - Answers***

*Database: Fuel Usage and Ship Emissions for 34 Ports in Australia*

*Link:* [*http://users.stat.ufl.edu/~winner/data/ship\_emissions2.csv*](http://users.stat.ufl.edu/~winner/data/ship_emissions2.csv)

Target variable (output) = fuel

Independent variables (input) = NO\_x, SO\_2, PM\_10

---------------------------------------------- RStudio -----------------------------------------------------

*data <- read.csv("http://users.stat.ufl.edu/~winner/data/ship\_emissions2.csv")*

*x1 =data$NO\_x # independent variable*

*x2 =data$SO\_2 # independent variable*

*x3 =data$PM\_10 # independent variable*

*y =data$Fuel # dependent variable*

*table(data$Fuel) # this function will help to identify the family*

*# data cleaning: removing missing values*

*dataset <- na.omit( data.frame(x1,x2,x3,y) )*

*# Fitting the Model*

*fit = glm (y~x1+x2+x3,data = dataset, family = 'gaussian') # linear regression as Fuel is a continuous variable*

*summary(fit)*

------------------------------------ RStudio - results -----------------------------------------------------

Call:

glm(formula = y ~ x1 + x2 + x3, family = "gaussian", data = dataset)

Deviance Residuals:

Min 1Q Median 3Q Max

-1519.5 -305.6 -3.1 350.6 2108.4

Coefficients:

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

(Intercept) -95.389 146.061 -0.653 0.519

x1 28.108 2.492 11.280 2.58e-12 \*\*\*

x2 108.721 8.790 12.368 2.61e-13 \*\*\*

x3 -990.235 93.423 -10.600 1.16e-11 \*\*\*

---

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

(Dispersion parameter for gaussian family taken to be 423911.7)

Null deviance: 1.2808e+10 on 33 degrees of freedom

Residual deviance: 1.2717e+07 on 30 degrees of freedom

AIC: 542.78

Number of Fisher Scoring iterations: 2

(b)

Parameters

β0 = -95.389

x1 = β1 = 28.108

x2 = β2 = 108.721

x3 = β3 = -990.235

β0 = p-value (0.519) > 0.05 (α), it means that this attribute is NOT significant and H0 (β0 = 0) hypothesis is accepted.

β1 = p-value (2.58e-12) < 0.05 (α), it means that this attribute is significant and H0 (β1 = 0) hypothesis is rejected.

β2 = p-value (2.61e-13) < 0.05 (α), it means that this attribute is significant and H0 (β2 = 0) hypothesis is rejected.

β3 = p-value (1.16e-11) < 0.05 (α), it means that this attribute is significant and H0 (β3 = 0) hypothesis is rejected.

(a)

---------------------------------------------- RStudio -----------------------------------------------------

*# split the dataset to trainset and testset*

*n=nrow(dataset)*

*indexes = sample(n, n\*(80/100))*

*trainset= dataset[indexes,]*

*testset = dataset[-indexes,]*

*# Fitting the Model*

*fit = glm (y~x1+x2+x3, data = trainset, family = 'gaussian') # linear regression as Fuel is a continuous variable*

*pred = predict(fit, testset)*

*pred*

*-------------------------------------------------------------------------------------------------------------*

(c)

------------------------------------ RStudio - results -----------------------------------------------------

> pred

3 6 12 20 29 30 32

60720.1786 23505.3927 8690.9964 2466.3409 1177.5519 299.4127 507.5632

Predictive model

Y = β0+ β1X1+ β2X2+ β3X3

Y = (28.108)X1 x (108.721)X2 x (-990.235)X3

(d)

---------------------------------------------- RStudio -----------------------------------------------------

*actual = testset$y # actual values for the target variable (y) to be used on the confusion matrix*

*pred\_fuel = ifelse(pred>=0.5,1,0)*

*pred\_fuel*

*table(actual)*

*table(pred\_fuel)*

*tab1 = table(pred\_fuel, actual) # create confusion matrix*

*tab1*

*accuracy = sum(tab1 [row(tab1) == col(tab1)])/ sum(tab1)*

*accuracy*

------------------------------------ RStudio - results -----------------------------------------------------

|  |
| --- |
| > tab1 = table(pred\_fuel, actual) # create confusion matrix  > tab1  actual  pred\_fuel 539 1645 1808 3147 3440 12499 15127  1 1 1 1 1 1 1 1  > accuracy = sum(tab1 [row(tab1) == col(tab1)])/ sum(tab1)  > accuracy  [1] 0.1428571 |
|  |
| |  | | --- | | > | |

**Question 2**

Let are identically independently distributed (iid) with Poisson().

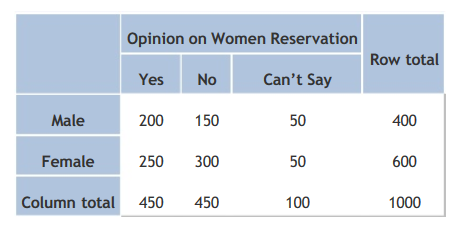
1. Compute the likelihood function (LF**). (10 Marks)**

1. Adopt the appropriate conjugate prior to the parameter (Hint: Choose hyperparameters optionally within the support of distribution). (**10 Marks)**
2. Using (a) and (b), find the posterior distribution of . **(10 Marks)**
3. Compute the minimum Bayesian risk estimator of . **(5 Marks)**

**(Total: 35 Marks)**

**Question 3**

An opinion poll surveyed a simple random sample of 1000 students. Respondents were classified by gender (male or female) and by opinion (Reservation for women, No Reservation, or No Opinion). Results are shown in the observed contingency table below.



Does the gender and opinion on women reservation are independent? Use a 0.05 level of significance. To do so,

1. State the hypotheses. **(5 Marks)**
2. Find the statistic and critical values. **(10 Marks)**
3. Explain your decision and Interpret results. **(15 Marks)**

**(Total: 30 Marks)**

***Question 3 - Answers***

(a)

Test of independence of two categorical variables

Let X1 be gender variable and X2 be opinion variable

H0: Variable X1 and Variable X2 are independent.

H1: Variable X1 and Variable X2 are not independent.

(b)

Statistic value = 16.20

Critical value = 5.99

---------------------------------------------- RStudio -----------------------------------------------------

*yes = c(male = 200, female = 250)*

*no = c(male = 150, female = 300)*

*csay = c(male = 50, female = 50)*

*o = data.frame(yes,no,csay)*

*t = chisq.test(o)*

*t\_value = t$statistic*

*t\_value # test-value*

*alpha = 0.05*

*df = t$parameter*

*df*

*c\_value = qchisq(1-alpha, df)*

*c\_value # critical-value*

------------------------------------ RStudio - results -----------------------------------------------------

> t\_value # test-value

X-squared

16.2037

> alpha = 0.05

> df = t$parameter

> df

df

2

> c\_value = qchisq(1-alpha, df)

> c\_value # critical-value

[1] 5.991465

(c)

Since teste-value is greater than critical-value we reject H0, it means that there is a relation between gender and opinion variables (they are NOT independent).