# 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.

<https://github.com/pdonada/B9DA101_S1-PD/tree/master/CA_TWO_(30%25)>

**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 **output** to **input** variables.

file\_read\_pd <- read.csv(file.choose()) #London2012Athletes.csv

summary(file\_read\_pd)

weight <- file\_read\_pd$Weight

height <- file\_read\_pd$Height..cm

dtframe <- data.frame(weight, height)

data\_pd <- na.omit(dtframe)

fit\_data\_pd <- glm(weight~height, data=data\_pd, family='gaussian') # linear regression

summary(fit\_data\_pd)

n <- nrow(data\_pd)

indexes <- sample(n, n\*(80/100))

trainset <- data\_pd[indexes,]

testset <- data\_pd[-indexes,]

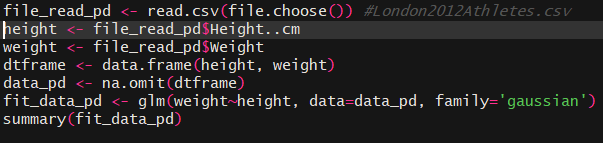
actual <- testset$weight

fit\_trainset <- glm(weight~height, data = trainset, family = 'gaussian')

summary(fit\_trainset)

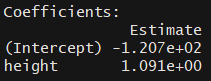
* 1. Specify the significant variables on the **output** variable at the level of 𝛼=0.05 and explore the related hypotheses test. Estimate the parameters of your model.

Estimating the parameters:



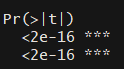
Beta 0: Intercept

Beta 1: height



Significance:

For Beta 0 and Beta 1 the P-value is **less** than Alfa (0.05) so both are significant variables:



Hypothesis test:

H0:B0 =0

H1: B0 <> 0

Since P-value < Alfa: Reject H0

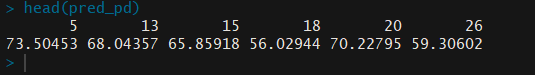
H0:B1 =0

H1: B1 <> 0

Since P-value < Alfa: Reject H0

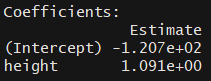
* 1. Predict the output of the test dataset using the trained model. Provide the functional form of the optimal predictive model.





In this example, we have only two parameters Beta0 and Beta1

Y = Beta0 x0 + Beta1 x1



Y = (-1.207e+02 \* x0 + 1.091e+00 \* x1)

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

file\_read\_pd <- read.csv(file.choose()) #London2012Athletes.csv

summary(file\_read\_pd)

weight <- file\_read\_pd$Weight

height <- file\_read\_pd$Height..cm

dtframe <- data.frame(weight, height)

data\_pd <- na.omit(dtframe)

fit\_data\_pd <- glm(weight~height, data=data\_pd, family='gaussian') # linear regression

summary(fit\_data\_pd)

n <- nrow(data\_pd)

indexes <- sample(n, n\*(80/100))

trainset <- data\_pd[indexes,]

testset <- data\_pd[-indexes,]

actual <- testset$weight

fit\_trainset <- glm(weight~height, data = trainset, family = 'gaussian')

summary(fit\_trainset)

pred\_pd <- predict(fit\_trainset, testset)

head(pred\_pd)

pred\_weight\_pd <- ifelse(pred\_pd >= 0.5, 1, 0)

head(pred\_weight\_pd)

table(actual)

table(pred\_weight\_pd)

tab1 = table(pred\_weight\_pd, actual) # create confusion matrix

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

accuracy



**Question 2**

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

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

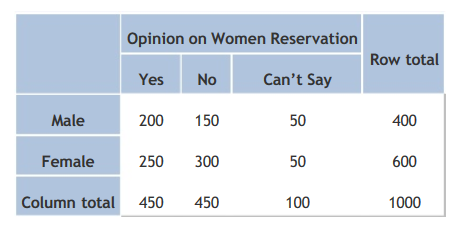
1. Adopt the appropriate conjugate prior to the parameter (Hint: Choose hyperparameters optionally within the support of distribution). (**10 Marks)**

1. Using (a) and (b), find the posterior distribution of . **(10 Marks)**
2. 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)**

Test of independent of two categorical variables (x1 and x2)

H0: x1 and x2 are independent

H1: Not H0

1. Find the statistic and critical values. **(10 Marks)**

α = 0.05

#name <- c('gender', 'yes', 'no', 'cant say')

male <- c(200, 150, 50)

fem <- c(250, 300, 50)

datafr <- data.frame(male,fem)

#datatb <- table(datafr$male, datafr$fem)

a <- chisq.test(datafr) # test value

a$statistic #testvalue

alpha <- 0.05

c\_value <- qchisq(1-alpha, a$parameter)

c\_value

statistic: 16.2037

critical values: 5.991465

1. Explain your decision and Interpret results. **(15 Marks)**

if test-value > c-value : R H0

since 16.2037 > 5.991465: Reject H0

so x1 and x2 are not independent

**(Total: 30 Marks)**