# Assessment brief

|  |  |
| --- | --- |
| Module Title: | Data Mining |
| Code: | B9DA103 |
| Module Leader: | Bahman Honari |
| Student code: | 10521647 |
| Student name: | Juliana Salvadori |
| Instruction for CA 2: | This CA if focused on two-stage Missing Data Imputation.    In the dataset provided, the first column is labelled as “Income” and is the column to be imputed. Other variables are the potential predictors in the models explained below.  The first step is to determine the missing values of “Income” that are actually equal to 0. To do this, you are required to define a new variable, say “Binary\_Income” with 0, 1 and missing values, for the cases which have 0, non-zero and missing values in the “Income” variable, respectively. Then use the cases with “Binary\_Income” equal to either 0 or 1, to develop a logistic regression model. The outcome of this model would allow you to predict the missing values in the “Binary\_Income” variable, as 0 or 1. For any case which “Binary\_Income” is predicted as 0, put the actual “Income” equal to 0 as well.  Those cases who their “Binary\_Income” is predicted as 1, are used along with other cases with nonzero “Income” to develop a linear regression model. The outcome of such model would allow you to predict the missing values for the “Income” variable, for the cases which their “Binary\_Income” variable is predicted as 1 and therefore are expected to have non-zero “Income”.  **Report**:  - The results for both logistic and linear regression models.  - The number of cases with predicted “Binary\_Income” as either 0 or 1.  - The Mean and SD for the non-zero predicted “Income”. |
| Deadline | 25 April 2020 |

## Logistic regression

Response variable = Binary\_Income

Independent variables = samptype, borough, sex, r\_age, educ, race, single, college, workmos, workhrs

Result =



Call:

glm(formula = Binary\_Income ~ ., family = "binomial", data = Data1)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.7932 0.0175 0.0282 0.0582 1.0239

Coefficients:

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

(Intercept) -0.37391 1.50676 -0.248 0.804015

samptype 0.24028 0.51986 0.462 0.643935

borough -0.35368 0.17186 -2.058 0.039597 \*

sex 0.16486 0.43489 0.379 0.704620

r\_age 0.04252 0.01125 3.778 0.000158 \*\*\*

educ 0.34584 0.30614 1.130 0.258610

race -0.09154 0.23594 -0.388 0.698025

single 0.10441 0.46079 0.227 0.820740

college -0.88612 0.81013 -1.094 0.274045

workmos 0.26643 0.16618 1.603 0.108862

workhrs 0.08806 0.04804 1.833 0.066773 .

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 311.46 on 1376 degrees of freedom

Residual deviance: 180.89 on 1366 degrees of freedom

AIC: 202.89

Number of Fisher Scoring iterations: 10



Accuracy of the model = 0.9789397

There were 124 rows where Binary\_Income was N/A and based on the logistic model above it was possible to assign 3 out of 124 as Binary\_Income = 0 and 121 rows as 1.

## Linear regression

Response variable = Income

Independent variables = samptype, borough, sex, r\_age, educ, race, single, college, workmos, workhrs

Result =



Call:

lm(formula = Income ~ ., data = Data2)

Residuals:

Min 1Q Median 3Q Max

-91668 -20746 -4807 12547 243719

Coefficients:

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

(Intercept) -15139.63 8320.49 -1.820 0.06905 .

samptype 2704.45 2113.66 1.280 0.20094

borough 1273.56 786.72 1.619 0.10572

sex -4224.71 1982.17 -2.131 0.03324 \*

r\_age 458.89 73.62 6.233 6.12e-10 \*\*\*

educ 10471.57 1516.19 6.907 7.67e-12 \*\*\*

race -5335.95 1029.20 -5.185 2.50e-07 \*\*\*

single -21834.78 2056.80 -10.616 < 2e-16 \*\*\*

college 10012.58 3266.91 3.065 0.00222 \*\*

workmos 1427.06 341.96 4.173 3.20e-05 \*\*\*

workhrs 584.08 79.62 7.335 3.82e-13 \*\*\*

---

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

Residual standard error: 34560 on 1333 degrees of freedom

(121 observations deleted due to missingness)

Multiple R-squared: 0.4387, Adjusted R-squared: 0.4345

F-statistic: 104.2 on 10 and 1333 DF, p-value: < 2.2e-16



## Mean / Standard Deviation

> # all data

> mean = mean(Data$Income[Data$Income > 0])

> mean

[1] 48915.19

> sd = sd(Data$Income[Data$Income > 0])

> sd

[1] 44852.05

> # only predicted rows

> mean = mean(Data$Income[Data$Income > 0 & Data$Dummy == 0])

> mean

[1] 54012.66

> sd = sd(Data$Income[Data$Income > 0 & Data$Dummy == 0])

> sd

[1] 27609.35