HOMEWORK 1 REPORT

Student id: 2020280082

Hong Massimo

Abstract

This report contains the experimental results obtained while trying to solve the problems given.

The main topics included in these experiments are:

- sampling
- regression methods

The assignment has been done utilizing Python as programming language, together with the following libraries:

- pandas: read the excel file containing the data and the simple random sampling
- sklearn: powerful library that has ready to use methods for linear and logistic regression
- matplotlib: plotting in exercise 1 part a
- numpy

The first exercise is divided into three major paragraphs, in which are described the respective settings, the obtained results and the interpretation of said outputs.

For the second question two sampling strategies have been used: simple random sampling method from the pandas library and an implementation of cluster sampling.

For the clustering sampling the data has been divided into 40 cluster, which means that each one has the size of 51, and in order to get the 10% of the original data size we need four clusters.

In the end there is a comparison between the results obtained by applying different sampling methods (including no sampling at all).

Exercise 1

Part a

Request: Consider the groups which have session number greater equal than 20 and predict the chatting behavior through the average age with linear regression. Calculate the mean and variance of the error.

After applying the constraint, the sample size has been reduced from 2040 to 1247 and the only columns we are interested in are: no response conversation ratio, night conversation ratio and image ratio and age variance.

For this particular problem, the average age represents the **x** and the other three columns are the labels we need to predict. This means that we need to perform three separate single linear regressions. In order to avoid too many repetitions and to keep the code cleaner, everything has been put inside a cycle that loops thrice. Each loop computes the linear regression for one feature.

Following are the parameters obtained by the linear regression.

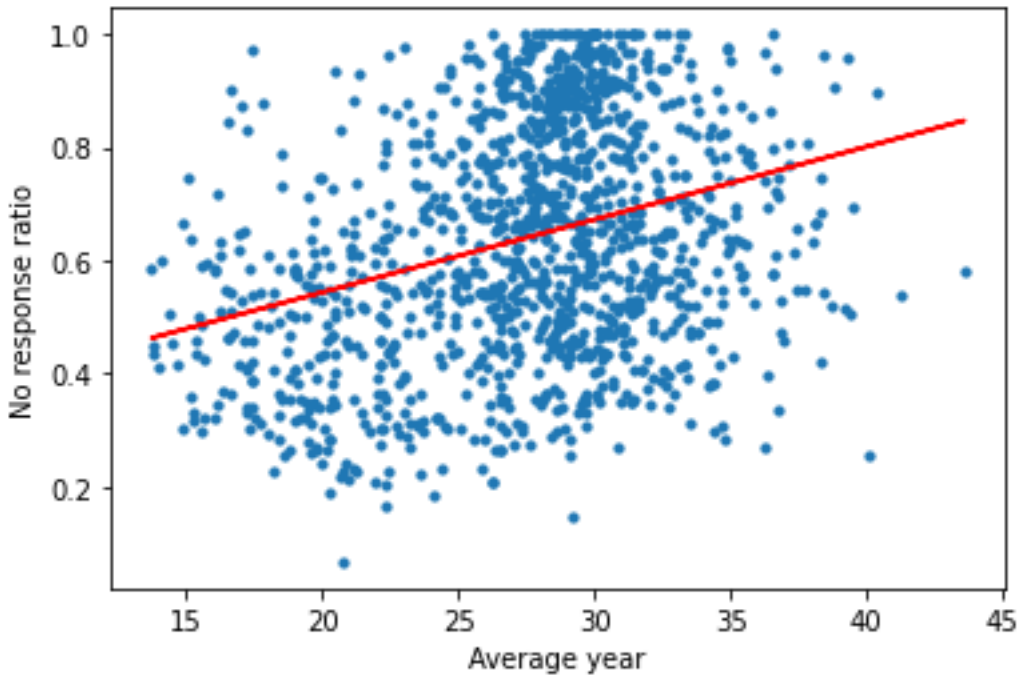
y = w0 + w1*x

w0 = intercept of the function

w1 = slope of the function

The correlation between average year and the respective features has been calculated using the numpy function **corrcoef**, which computes the Pearson correlation coefficient.

Predict no response conversation ratio through average age



w0 = 0.2862932341884706

w1 = [0.01287497]

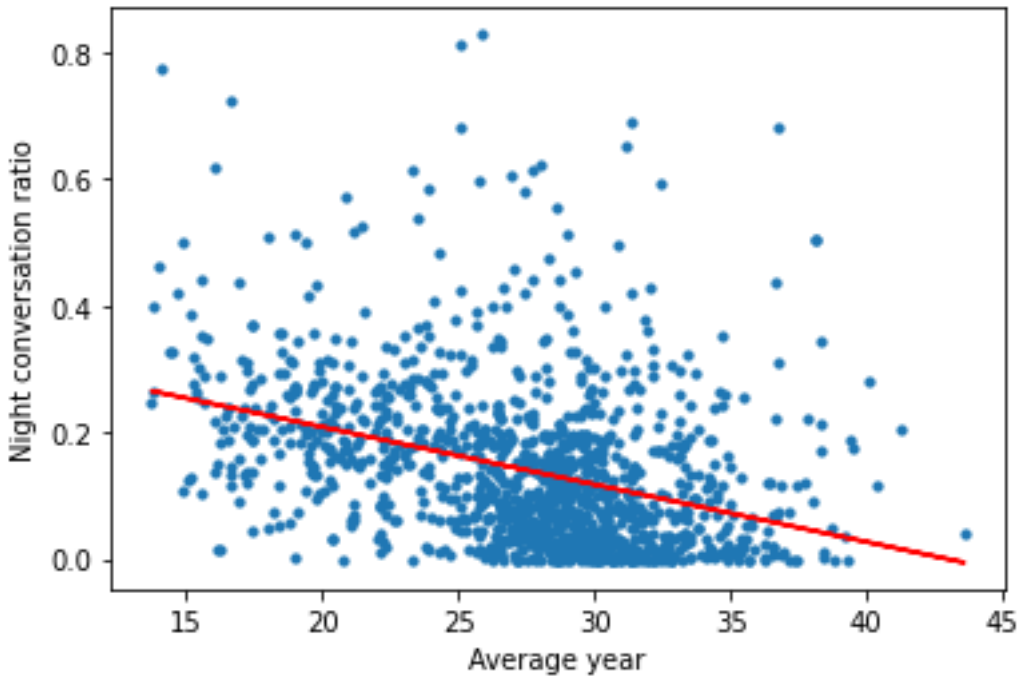
Error variance = 0.01128521210432624

The correlation between x and y is:

[[1.   0.31049854]

[0.31049854 1.   ]]

Predict night conversation ratio through average age



w0 = 0.38992886214263767

w1 = [-0.00907208]

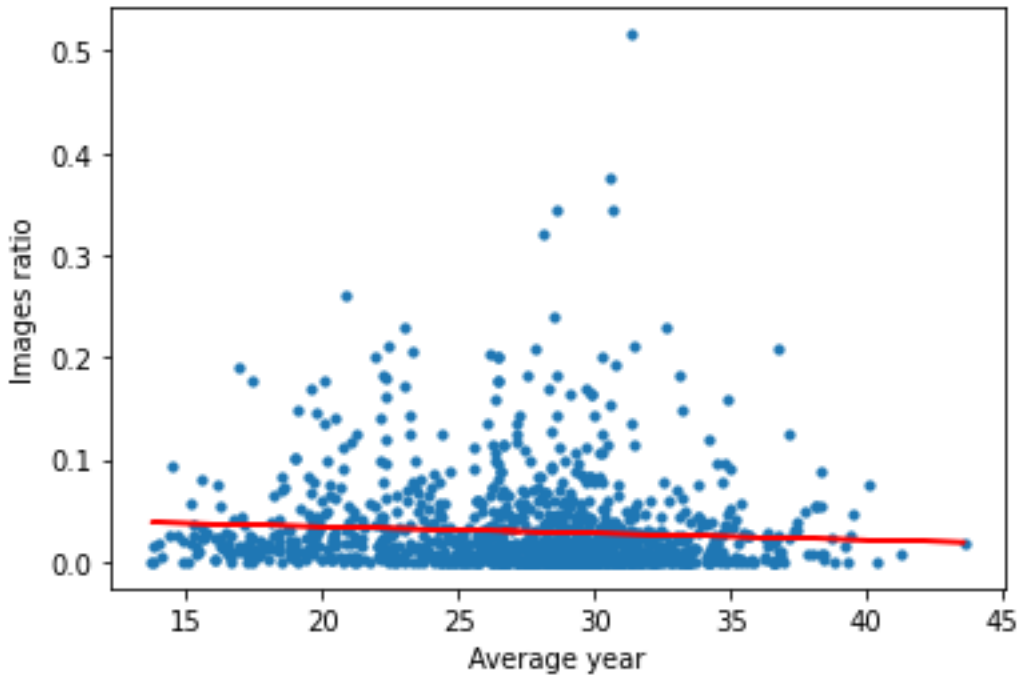
Error variance = 0.006580046205249687

The correlation between x and y is:

[[1.   -0.36276324]

[-0.36276324 1.   ]]

Predict image ratio through average age



w0 = 0.04738408786102775

w1 = [-0.00065006]

Error variance = 0.0012780259377178952

The correlation between x and y is:

[[1. -0.07134]

[-0.07134 1.]]

Part b

Request: perform weighted multivariate linear regression to predict no response conversation ratio, night conversation ratio and image ratio through other features and using session numbers as weights.

No response conversation ratio

w0 = 0.47997901293427003

w1 = [-5.38390841e-05, -4.68122627e-06, -4.12202947e-02, -2.13593440e-02

7.94852627e-03, -2.49244576e-03, -4.24498335e-02, -1.85867365e-01]

Error variance = 0.02182273435813518

Night conversation ratio

w0 = 0.2804630956166998

w1 = [-4.68507762e-, 3.54148670e-07, 4.53317569e-02, 3.92147641e-02

 -8.38068734e-03, 1.05094105e-02, -4.25586808e-02, 1.30062681e-01]

Error variance = 0.01899818588150442

Images ratio

w0 = 0.043845483651511676

w1 = [-2.70603813e-05, 6.03498547e-07, 2.52727446e-04, 2.36349628e-02,

 -7.81753362e-04, 4.88163181e-04, -1.17837768e-02, -8.52112529e-03]

Error variance = 0.005154034567311918

Part c

1)

Request: Filter groups belonging to category 1 and 4 and divide the population into train set and valid set. Classify the set through logistic regression using other features.

For this part I have chosen the features that I personally thought were the most important, and therefore, able to better predict the category.

The chosen features are: sex ratio, average age, session number, no response ratio, night conversation ratio, images ratio.

The model has been fitted with the train set, while its accuracy has been judged by scoring the test set. The population has been divided by utilizing the train_test_split() function offered by sklearn.

Below are the results obtained.

w0 =[-1.12555062]

w1 = [[-2.25053796 , 0.14322032, -0.00582243, -1.16789569, 0.20764413, 0.13072837]]

Score = 0.6973684210526315

2)

For the second portion of the exercise I have eliminated one feature at a time to try to find the most important feature for a better prediction.

- The image ratio feature has been eliminated

  w0 = [-1.12399587]

  w1 = [[-2.24530914 0.14332052 -0.00582446 -1.17304887 0.21221977]]

  score = 0.6885964912280702

- The night conversation ratio has been eliminated

  w0 = [-1.09586732]

  w1= [[-2.23966193 0.14262638 -0.00579033 -1.14325704 0.09819777]]

score = 0.6973684210526315

- The no response ratio has been eliminated

  w0 = [-1.76192696]

  w1 = [[-2.3531012 0.13601163 -0.0045442 0.18238697 0.19321069]]

  score = 0.6885964912280702

- The session number has been eliminated

  w0 = [-2.09430285]

  w1 = [[-2.73710602 0.14388583 0.08157791 0.17023012 -0.04698181]]

  score = 0.6535087719298246

- The average age has been eliminated

  w0 = [1.78426092]

  w1 = [[-1.55618936 -0.00573354 -0.77825585 -0.16186949 0.51585085]]

  score = 0.6929824561403509

- The sex ratio has been eliminated

  w0 = [-1.84628492]

  w1 = [[0.12252796 -0.00659739 -1.35522037 0.22709592 0.18243899]]

  score = 0.6535087719298246

We can see that the sex ratio and the session number are the most important features between the chosen ones, as the accuracy of the predictions lower much more when they are eliminated.

3)

Request: multi-class classification on all categories

The settings of this experiment are the exact same as for part 1.

Here are the results obtained.

w0 = [5.12874356 -2.85453476 -2.88036321 3.69990846 -3.09375405]

w1 = [[1.71736341e+00 -2.31178354e-01 1.20730304e-03 -8.02919736e-02
7.62401419e-01 6.99076684e-01]
[-3.01656233e-01 9.55772052e-02 1.73821774e-03 4.29632750e-01
-9.21664288e-01 -6.61341277e-02]
[-2.29970316e+00 1.51117511e-01 8.00406498e-04 -8.86476038e-01
6.79182333e-01 -9.27226091e-01]
[-1.12971161e+00 -4.98220490e-02 -6.28283003e-03 -1.48106176e+00
9.63398797e-01 1.20080916e+00]

[2.01370760e+00 3.43056865e-02 2.53690276e-03 2.01819702e+00

-1.48331826e+00 -9.06525628e-01]]

score = 0.4745098039215686

It is very obvious that the accuracy of the logistic regression is much lower when trying to predict non binary classes.

Exercise 2

Request: use linear regression to predict the chatting behavior based of average age using different sampling methods.

Sampling strategies implemented:

- Simple random sampling – using the default function sample() offered by pandas
- Cluster sampling – the population has been divided into 40 clusters of size 51, so in order to get 10% of the data we need 4 clusters.

Below are the results gotten from the experiment after repeating each sampling strategy 1000 times.

No-response conversation ratio

Random simple sampling

mean of intercepts = 0.45609719999087267

mean of slopes = 0.009215423930469376

variance of simple intercepts 0.007965879839517173

variance of simple slopes 1.0026223432668331e-05

variance of simple error variance 0.05576138785817069

Clustering

mean of intercepts = 0.5355641646924529

mean of slopes = 0.00013829772849727057

Variance of intercepts = 0.7362433144855011

variance of slopes = 4.395219008543582e-06

variance of cluster error variance 0.055916362932914083

Night conversation ratio

Random sampling

mean of intercepts = 0.36073376907965915

mean of slopes = -0.007901885237258079

variance of intercepts = 0.006022541916305189

variance of slopes =7.423785618556051e-06

variance of simple error variance 0.0348560327087401

Clustering

mean of intercepts 0.30700896485911716

mean of slopes -0.0001220925914346981

Variance of cluster intercepts 0.42715939257539853

variance of cluster slopes 2.5057294838071596e-06

variance of cluster error variance 0.03481308105030429

Images ratio

Random sampling

mean of simple intercepts 0.049272117314909256

mean of simple slopes -0.000615599008589287

variance of simple intercepts 0.0010200511127297144

variance of simple slopes 1.187933307244839e-06

variance of simple error variance 0.006603225036549536

Clustering

mean of cluster intercepts 0.0770012145419462

mean of cluster slopes -3.284999722932818e-05

Variance of cluster intercepts 0.07222599316416496

variance of cluster slopes 4.761069325414986e-07

variance of cluster error variance 0.006637165108105502

From the above data we can observe that sometimes cluster has a lower variance, but in the majority of the cases simple random sampling is slightly better.

After comparing the error variances between sampling and without sampling we can confirm that generally speaking, the error variance without sampling is lower that with sampling. This might be caused by the fact that without sampling we have a bigger sample size to work on.