Homework 6

Jichao Yang

Problem 1

(a) The dataset is prepared by dropping entries with empy data, normalizing the `age` variable, encoding the `passengerClass` variable with dummies, and adding an interept constant column.

```
import pandas as pd

# Read the dataset
df = pd.read_csv('../../data/TitanicSurvival.csv', index_col='rownames')
df = df.reset_index(drop=True)
df = df.dropna()

# Encode categorical variables
df['survived'] = df['survived'] == 'yes'
df['is_male'] = df['sex'] == 'male'
# We do not encode 3rd class to avoid colinearity
df['is_1'] = df['passengerClass'] == '1st'
df['is_2'] = df['passengerClass'] == '2nd'
# Normalize age variable
df['age'] = (df['age']-df['age'].mean())/df['age'].std()
# Change var type to float for training
df = df.drop(columns=['sex', 'passengerClass'])
df = df.astype('float')

df.describe().round(3)
```

	survived	age	is_male	is_1	is_2
count	1046.000	1046.000	1046.000	1046.000	1046.000
mean	0.408	0.000	0.629	0.272	0.250
std	0.492	1.000	0.483	0.445	0.433
min	0.000	-2.062	0.000	0.000	0.000
25%	0.000	-0.616	0.000	0.000	0.000
50%	0.000	-0.131	1.000	0.000	0.000
75%	1.000	0.633	1.000	1.000	0.000
max	1.000	3.477	1.000	1.000	1.000

(b) Recall that the 29 year old female passenger in first class can be found in `df.iloc[0]`. The implementation can be found as follow:

```
df.iloc[0]
```

```
survived              yes
sex                female
age                  29.0
passengerClass        1st
Name: 0, dtype: object
```