Using Machine Learning to Build a Scalable Tool to support Dieticians to Fight Chronic Diseases

Moumita Bhattacharya and Debarati Roychowdhury

MOTIVATION

- We propose to assist dieticians by providing information such as what factors are most indicative of *Fat* or *Protein* intake; what are the differentiating factors among people who consume more *Protein* compared to those who consume more *Fat*.
- We plan to utilize Machine Learning techniques such as Feature Selection to identify most informative factors for nutrient intakes.

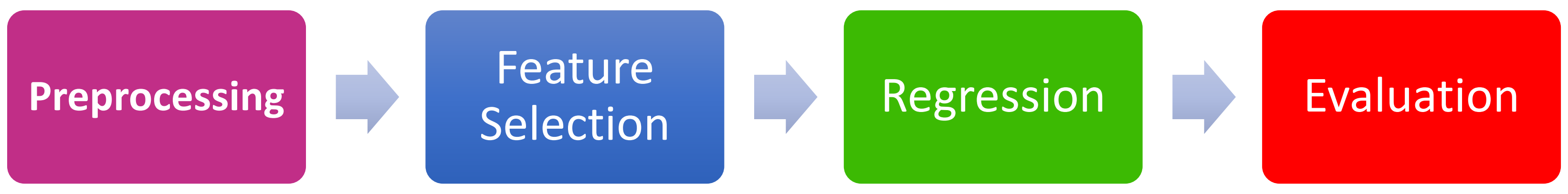
RESEARCH QUESTIONS

- *Which demographic attributes are most informative for the predicting the macro nutrients?*
- *Does nutrient intake profile patterns differ by fields such as gender, age, race, education, poverty level?*

GOAL

- We aim to develop a framework using **Big Data** and **Machine Learning** models that can assist dieticians to identify predictive patterns in nutrient intake, which can then help in fighting chronic diseases such as Obesity.
- We hypothesis that our approach can be extended to be used by dietician in real-time to provide nutrient intake advice.

METHODOLOGY



Preprocessing → Feature Selection → Regression → Evaluation

Preprocessing

- We remove four attributes that have a lot of missing values as well as the ones that have information only pertaining to children and old individuals.
- We then use *MapReduce in Spark* to obtain the sum of values of each of the five macro-nutrients for each individual and average the macro-nutrient intake values over two days.

Feature Selection

- The process of extracting the most informative attributes from the dataset
- We use a ***Regularized Linear Regression*** for feature selection.
- Specifically, we use a combination of ***LASSO*** and ***RIDGE*** regression - ***ElasticNet***

Regression

- ***Linear Regression*** *and* ***Regularized Linear Regression***

Evaluation

- ***Empirical validation***

BACKGROUND

- The ***NHANES*** is a nationwide survey conducted by the National Center for Health Statistics and some other health agencies since 1971 [1].
- The aims of the survey is to provide nationally representative information on ***nutritional status*** of the ***population*** and ***tracking*** changes over ***time***.
- ***Macro Nutrients*** *– Fat, Carbohydrate, Protein, Fiber, Sugar*
- ***Individual Record*** *– Demographic Attributes e.g. Gender, Race, Income*

DATASET

- We obtain the demographic details and the nutrient intake record-sets of the individuals from the NHANES website [1].
- Example **record set** of three individual for Fat intake:

Individual 1	*< ID1, Fat, Age, Gender, Education Level, Income,….,>*
Individual 2	*<ID2, Fat, Age, Gender, Education Level, Income,…., >*
Individual N	*<IDN, Fat, Age, Gender, Education Level, Income,…., >*

RESULTS

TABLE 1. shows that Fat intake varies significantly with age however the variation is not significant for gender

	FAT	CARB	PROTEIN	FIBER	SUGAR
AGE	X	X		X	X
GENDER			X		X
RACE			X	X	
INCOME	X	X	X	X	X
EDUCATION LEVEL	X	X	X	X	
COUNTRY OF BIRTH	X	X			X
SPOKEN LANGUAGE		X	X	X	X
NO. OF PEOPLE IN HH		X	X	X	X
PREGNANCY STATUS	X	X			X
INTERPRETER USED	X	X			X

RESULTS

Fig. 1 and Fig. 2 shows that Fat intake varies significantly with age however the variation is not significant for gender

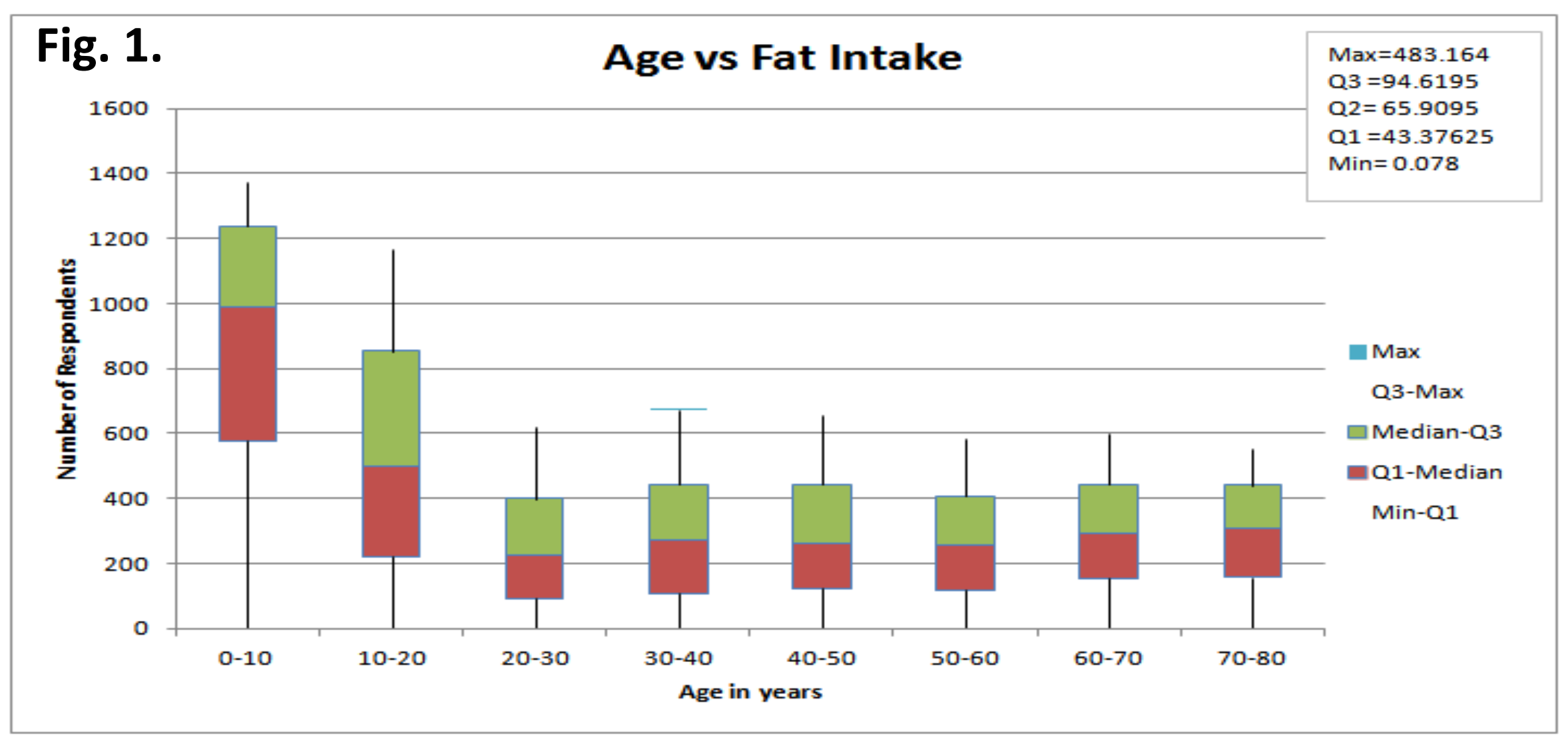


Fig. 1.

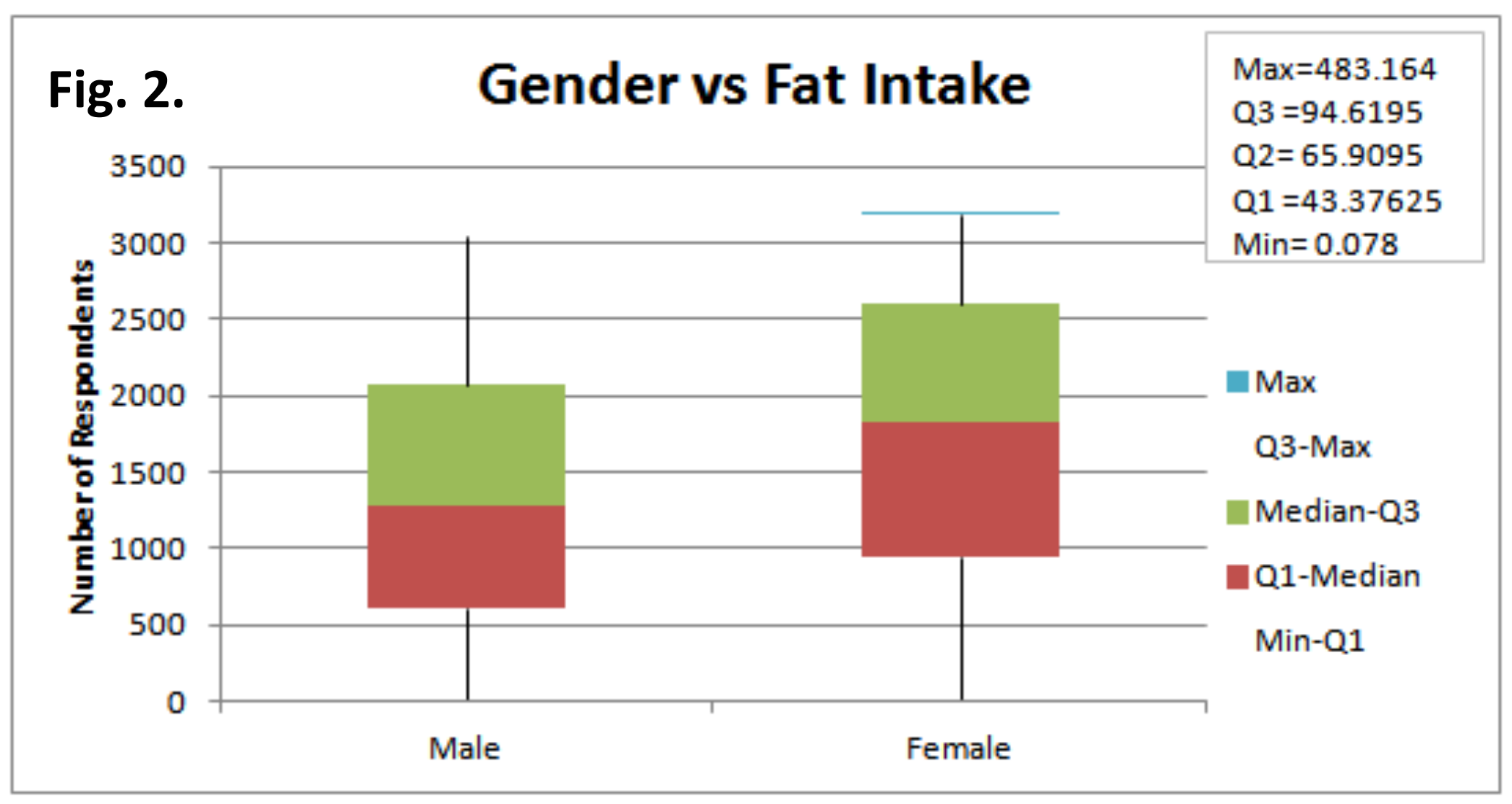


Fig. 2.

Fig. 3 shows variation of macro nutrient intake with age

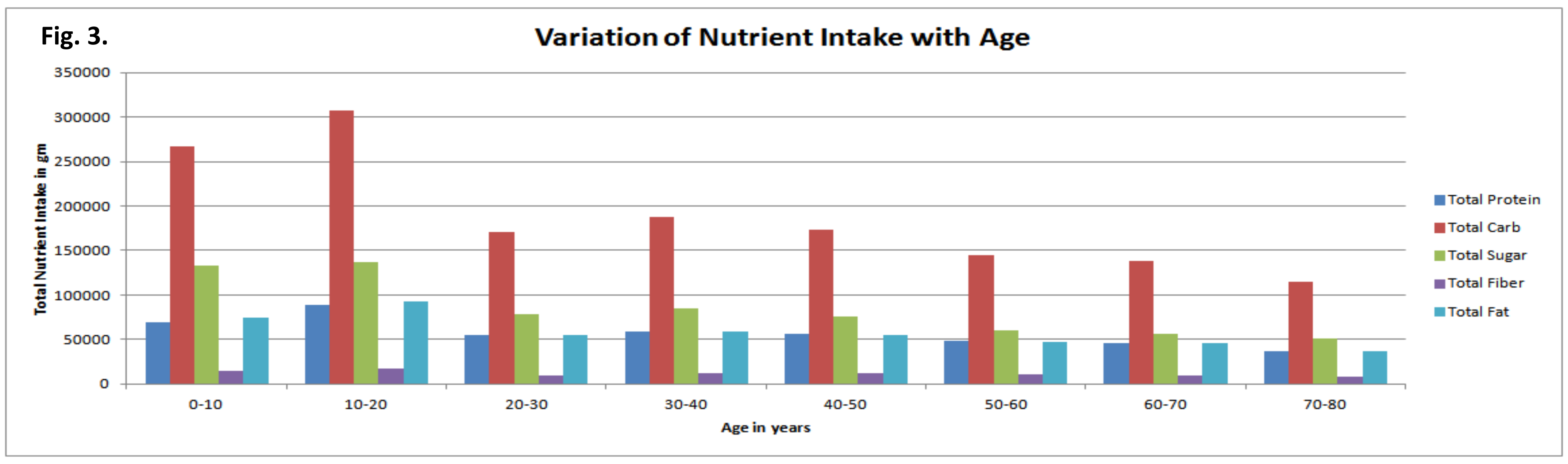


Fig. 3.

Fig. 4 shows variation of macro nutrient intake with Income

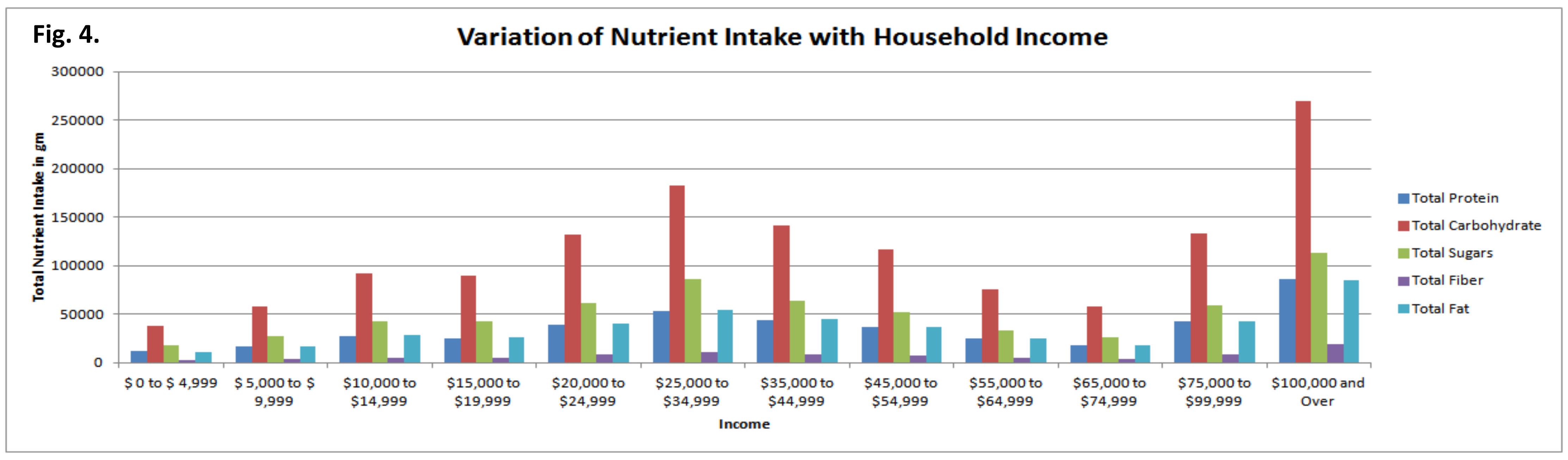


Fig. 4.

DISCUSSION

- Number of features that are highly informative of *Carbohydrates* and *Sugar* is significantly more than the number of informative features of *Fat*, *Protein* and *Fiber*, indicating that the latter macro nutrients are impacted by smaller number of demographic features compared to the former.
- Our results show that *Gender* is a highly informative feature for *Protein* and *Sugar* intake but not a clear indicator of *Fat*, *Carbohydrate* and *Fiber* intake.
- We empirically validate the above observation by plotting *Fat* intake for different *Gender* and *Age ranges* (See Figures 1 and 2) and observe that *Fat* intake indeed does not vary with respect to *Gender*. However, *Fat* intake significantly varies among different *Age ranges*. Furthermore, we intent to validate this result by surveying the existing literature.
- Table 1 shows that *Age* and *Income Level* are highly informative of most of the macro-nutrients. We demonstrate this by plotting bar graphs (See Figures 3 and 4) representing the variation of all macro-nutrients intake with respect to different *Age ranges* as well as *Income Levels*.

FUTURE WORK

- We plan to implement the feature selection and regression models using the datasets from all the years available in the NHANES website.
- We also plan to utilize clustering methods to identify individuals whose macro-nutrient intake profile is similar, enabling evaluation of the existing results.

REFERENCES

[1] Centers for Disease Control and Prevention (2016). Nation Health and Nutrition Examination Survey. *https://www.cdc.gov/nchs/nhanes/*, last accessed 12/05/16.