JSC270H1: Assignment 2 Report Xiang Gao

Background. Linear regression is a fundamental and crucial statistical machine learning method that is widely employed; its simplicity and efficiency allow people to obtain excellent models in certain settings. Using this powerful tool, we will explore a supervised learning setting with the US census data extracted by Kohavi and Becker in 1995. Specifically, we will focus on learning how weekly working hours of people linearly relates to their gender, education background, as well as the income level.

Motivation. Studies on census data mostly showed that men earned more money per year than women on average; however, working hours are not usually considered as a factor for the study. Do men work more than women? There are other confounders for this study that we will no be focusing on: women spend more time doing unpaid work such as housekeeping and taking care of the children. Instead, other factors such as education and actual earnings will be taken in account.

Methods. In this section, we describe the methodology for modelling the relationship between weekly hours and other factors. Estimates for the parameters of the model are obtained using ordinary least squares. Also, we make inference with the regression coefficients through hypothesis testing; the null hypothesis is the coefficients are 0 which means that there are no linear association among the variables. Various statistics such as AIC (Akaike Information Criterion) are calculated to assess the models being fitted. We assessed three models each with different features. The computation is completed through *ols()* method from *statsmodels.formula.api*.

Results. The fitted equation for the first model with feature gender is $\hat{y} = 6.0177x + 36.4164$ where the output is men's working hours when x is equal to 1. The p-value for

slope parameter is 0.000, and an AIC of 2.543e+05. The second fitted model is based on gender and the number of years of education: $\hat{y} = 29.4106 + 5.9709x_1 + 0.6975x_2$ where the output is men's working hours when x_1 is equal to 1. The p-value for slope parameter is 0.000, and an AIC of 2.536e+05. The third fitted model is based on gender and the number of years of education: $\hat{y} = 31.4218 + 5.1010x_1 + 0.4478x_2 + 4.5175x_3$ where the output is working hours of men with income over 50K when x_1 and x_3 is equal to 1. The p-value for slope parameter is 0.000, and an AIC of 2.529e+05.

Discussion. From the fitted models, they suggest that men work at least 5 hours more than women weekly, and the people who earns more than 50K works 4.5 hours more than the people who earn less than 50K. Moreover, the education time length also has a positive association. The p-values of all the regression coefficients in each model show a strong evidence that we can reject the null hypothesis that there are no linear association between the variables. The lowest AIC is achieved in the model with three parameters. Since AIC itself already considers the problem of overfitting, we are safe to go with model three without being worried about generalization and overparameterization.

Conclusion. In this study, we used linear regression to study how a person's various features linearly associate with their weekly working hours and made inference with the regression coefficients using hypothesis testing. The obtained p-values support our fitted models. With the AIC information, the most complex model with three parameters in our study is the best one among 3. The results indicate that men work more than women weekly, and people earning more money also works more. However, there are some confounders in this study that we ignored; hence, we cannot conclude further than what we have studied.