JSC270 Winter 2024 Assignment 2

Part III Reporting on my own Regression Analysis

Description of Dataset

This dataset was cleaned and extracted from the 1994 US Census by Kohavi and Becker in 1995. A detailed description of it can be found here. Originally intended to predict whether a person makes over $50K a year, it is now extensively used in various studies to understand socio-economic factors affecting income level, employment patterns, etc. It serves as a benchmark dataset for papers advancing statistical and machine learning methods. The dataset contains 32561 observations and 15 variables. These variables include *age*, *workclass*, *fnlwgt*, *education*, *education_num*, *marital_status*, *occupation*, *relationship*, *race*, *sex*, *capital_gain*, *capital_loss*, *hours_per_week*, *native_country*, and *gross_income_group*. Continuous variables include *age*, *fnlwgt*, *education_num*, *capital_gain*, *capital_loss*, and *capital_loss*. The rest are categorical variables. Despite minor missing values in *workclass*, *occupation*, and *native_country*, the dataset is largely complete.

The Question

Amidst the demographic shifts in the US population in 1994, I wanted to investigate how race might influence the number of hours worked per week. Using a linear regression model, I wanted to answer the question: *During the year of 1994, did the average number of hours worked per week vary for people of different races?* This investigation can help further our understanding how societal factors such as race intersect with the economy and labour market in the past.

Model Description

There were only two variables I included in the model: *race*, as the independent variable, and *hours_per_week*, as the dependent variable. The *race* variable was a categorical variable containing 5 possible values: *White*, *Asian-Pac-Islander*, *Amer-Indian-Eskimo*, *Other*, *Black*. The *hours_per_week* variable is a continuous variable indicating how many hours a person works each week. It ranges from 40 to 94. These two variables have a direct relevance towards the original question. And by focusing on just these two variables, the analysis will be able to provide a simple yet insightful perspective on the relationship between race and work hours. It is

important to note that these two variables do not have any missing variables, so we don't need to fix or replace anything.

Interpretation and Description of Results

Using the *smf.ols* and *fit* function, we get a least squares regression model with *race* as an independent variable and *hours_per_week* as a dependent variable. Taking a look at the coefficients, we have the intercept coefficient as 40.0482, *race[T.Asian-Pac-Islander]* coefficient as 0.0788, *race[T.Black]* coefficient as -1.6254, *race[T.Other]* coefficient as -0.5796, and *race[T.White]* coefficient as 0.6409. The coefficients represent the mean difference of number of hours worked between that group and the intercept. Since *Amer-Indian-Eskimo* is not included, we know that the intercept represents the mean hours worked per week for *Amer-Indian-Eskimo*. We see that the people who identified as Black had a notable lower mean number of hours worked compared to the rest of the groups. Taking a look at the p-values for the coefficients, we see that the intercept and *race[T.Black]* were the only ones with a p-value less than $\alpha=0.05$, indicating only those two were statistically significant. Looking at the 95% confidence interval, we see that for Black people, it falls within a negative range of [-3.062, -0.188], indicating that if we took many samples and built a regression model for each, the coefficient will fall within this range 95% of the time. These results seem to indicate that Black people worked less hours compared to the rest of the groups. A likely explanation for this is the US's long history of discrimination towards members of minority groups and especially the Black community. However, although statistically significant, it is also important to note the coefficient of determination is 0.003, indicating a weak association. This suggests a need for further exploration incorporating broader demographic factors.

In conclusion, the analysis sheds light onto the relationship between race and the number of hours worked in US during 1994. For future investigation, it is important to consider other models accounting for any possible confounding variables, such as age or relationship status, and the potential database biases like uneven distribution of races within the census. We can compare future models with this one using either the coefficient of determination or the residual squared error (of about 138.09).