Regression Analysis Report

Runshi Yang

The dataset I used for this assignment is called “Adult Data Set” (also known as “Census Income” Dataset), which is extracted and cleaned from the 1994 Census database by Barry Becker. The dataset has 15 columns and 32561 rows, and it was used to predict whether a person makes over 50K dollars a year.

As a student, I am interested in the question of whether black people and white people of the same age had equal access to educational resources at the time. The race of each observation in this dataset may be ‘Black’, ‘White’, ‘Asian-Pac-Islander’, ‘Amer-Indian-Eskimo’ and ‘Other’. So, I filtered out the rows with ‘Black’ and ‘White’ races before fitting the linear regression model. In addition to the ‘race’ variable, I also included ‘age’. This is because in Correlation.1.c of PART II, we learned that age is an important factor affecting educational attainment, so to simply study the effect of race on education, we need to control for a constant age.

Here's a summary of the linear regression model I came up with：

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.2257	0.061	152.475	0.000	9.107	9.344
race[T. White]	0.6421	0.048	13.381	0.000	0.548	0.736
age	0.0069	0.001	6.529	0.000	0.005	0.009

The intercept (9.2257) is the expected level of education of a 0-year-old black person, which is meaningless, since this it is absurd to consider the level of education of a newborn child. Since we are only interested in the relationship between race and educational attainment, I will only explain the coefficient corresponding to race. The coefficient for race is the estimated expected difference in education level between white people and black

people, holding age constant. So, we can see that a white person receives on average 0.6421 higher levels of education than a black person of the same age. The last two columns give us more information, we are 95% sure that a white person receives on average from 0.548 to 0.736 higher levels of education than a black person of the same age. And notice that all the p-values in the fourth column are very small (p-value < 0.05), we have strong evidence against the null hypothesis that the slope parameter of race is equal to 0, which means we were able to determine to a large extent that black people had less access to the same educational resources as white people of the same age at the time. However, the R-squared and adjusted R-squared are all 0.007, which means only 0.7% of the data fit the model well, so this is not a very good model.