**Assignment #2**

**20175404 Rocku Oh**

***Dataset***

The dataset, “The Adult Dataset”, utilized in this assignment is about thirty thousand records with 15 variables consisting six continuous (age, fnlwgt, education year, cap gain, cap loss, and hours per week) two binomial (salary and sex), six categorical (work class, occupation, relationship, native country, race, and marital), and one ordinal (education) variable. The data has been cleaned by removing each row when the missing value exists at the workclass, occupation, and native-country variables. After cleaning, the total records is 30162.

***Problems***

1. **Predict the age of a female member of the labor force by looking at the**
2. ***number of hours she works in a week,***
3. ***her marital status***
4. ***her education year***
5. ***her race***
6. ***her salary***

In the case of the female data, marital status can be re-categorized as a reference category, “Never-married”, because the category has largest data records. The “White” category in “Race” variable have chosen as a reference category based on the number of data record as shown in Table 1. The information of the re-categorized dummy variable are described in Table 2.

Based on the coded dummy variables, regression analysis is conducted. As shown in Table 3 Column (1), The R2 is 0.303 and F-statistic indicate the regression model is statistically valid (significance level of .01).

Table 1

*Bi-variate (“Marital” and “Race”) Frequency Table for “Female”*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Race | | | | | |
| Marital status | Amer-Indian-Eskimo | Asian-Pac-Islander | Black | Other | White | All |
| Never-married | 34 | 137 | 648 | 38 | 3455 | **4312** |
| Married-AF-spouse | 0 | 0 | 1 | 0 | 11 | 12 |
| Married-civ-spouse | 19 | 69 | 138 | 14 | 1240 | 1480 |
| Married-spouse-absent | 5 | 17 | 37 | 7 | 123 | 189 |
| Divorced | 33 | 44 | 321 | 16 | 2115 | 2529 |
| Separated | 7 | 10 | 162 | 9 | 386 | 574 |
| Widowed | 9 | 17 | 92 | 3 | 565 | 686 |
| All | 107 | 294 | 1399 | 87 | **7895** | 9782 |

Table 2

*Dummy variables Codification of Categorical Variables*

|  |  |  |
| --- | --- | --- |
| Dummy variable | Code | Initial variable |
| *Marital status* |  |  |
| Have-Married | 1 | Divorced, Separated, Widowed, Married-AF-spouse,  Married-civ-spouse, Married-spouse-absent |
| Never-married | 0 | Never-married |
| *Sex* |  |  |
| Male | 1 | Male |
| Female | 0 | Female |
| *Race* |  |  |
| White | 1 | White |
| Others | 0 | Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other |
| *Salary* |  |  |
| >50K | 1 | >50K |
| <=50K | 0 | <=50K |

Table 3

*Ordinary Least-square of Age Estimation with the Adult Census Data*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | OLS  (1) | OLS  (2) | OLS  (3) | OLS  (4) | OLS  (5) |
| *Dep. Variable* | Age | Age | log(Age) | log(Age) | log(Age) |
| *Regressors* |  |  |  |  |  |
| HoursPWeek | 0.0451\*\*\*  (4.382) | -0.0235\*\*\*  (-4.132) | 0.0027\*\*\*  (18.803) | 0.0026\*\*\*  (18.655) | 0.0026\*\*\*  (18.65) |
| Marital  (0=Never-married) | 14.7463\*\*\*  (61.688) | 14.1582\*\*\*  (95.843) | 0.4139\*\*\*  (119.523) | 0.4139\*\*\*  (119.531) | 0.414\*\*\*  (119.555) |
| Race  (1=White) | -0.7854\*\*\*  (-2.700) | -0.4215\*\*\*  (-2.263) | -0.0299\*\*\*  (-6.775) | -0.0300\*\*\*  (-6.790) | -0.0342\*\*\*  (-7.075) |
| EducationYr | 0.0152  (0.296) | 0.0106  (0.394) | 0.0061\*\*\*  (9.43) | 0.0060\*\*\*  (9.159) | 0.0060\*\*\*  (9.173) |
| Salary  (1= “>50K”) | 0.6655\*  (1.732) | 2.7400\*\*\*  (16.077) | 0.0710\*\*\*  (17.876) | 0.0675\*\*\*  (16.641) | 0.0445\*\*\*  (3.876) |
| Sex  (1=Male) | - | -0.5563\*\*\*  (-3.851) | -0.0214\*\*\*  (-6.319) | -0.0214\*\*\*  (-6.336) | -0.0213\*\*\*  (-6.300) |
| CapGain | - | - | - | 7.259E-07\*\*\*  (3.520) | 7.340E-07\*\*\*  (3.542) |
| CapLoss | - | - | - | 8.838E-06\*\*  (2.345) | 8.816E-06\*\*  (2.339) |
| Salary \* Race | - | - | - | - | 0.0253\*\*  (2.146) |
| Constant | 27.3771\*\*\*  (42.900) | 29.755\*\*\*  (80.111) | 3.1371\*\*\* (349.285) | 3.1391\*\*\*  (349.085) | 3.1425\*\*\*  (344.122) |
| Observations | 9782 | 30162 | 30162 | 30162 | 30162 |
| R-squared | 0.303 | 0.282 | 0.423 | 0.424 | 0.345 |
| F-statistic | 851.2 | 1972.0 | 3349.0 | 2628. | 1763. |

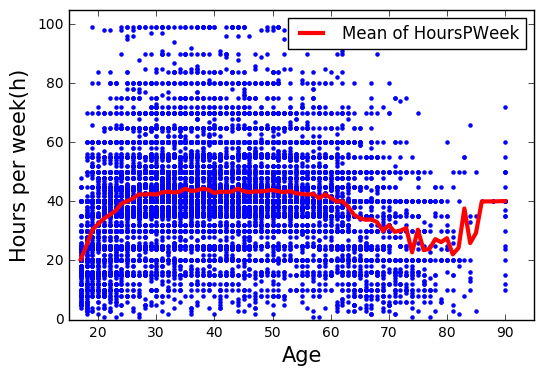
*Note.* Column (1) and (2) uses the real values of age. Column (3), (4), and (5) uses the log scale of the age as dependent variable. Especially, interaction term between salary and race is included at the model of column (5). The t-statistics are given in the parentheses with starts indicating \*\*\* p < 0.01, \*\* p < 0.05, \* p < 0.1.

1. **Interpret the coefficients.**

As shown in the Table 2 Column (1), the age difference between never-married and have-married female is about 14.7 year which is the biggest effect to determine the dependent variable. In the case of the race factor, even if the factor is statistically significant (significance level of 0.01), the impact on age is not very much (-0.7854). Education year and salary factor cannot reject null hypothesis because of the small t-statistic. However, predicted age increased as the education year increased. A female who earned salary of larger than 50K is 0.66 year older than the others.

1. **What is the predicted age of your instructor? He works 168 hours every week, is not married, has been educated for 19 years, is Asian, and gets paid less than 50K.**

The predicted age of the female is about 35 (35.241). However, the input value of 168 hours/week are not realistic. The maximum value of the hours per week is 100 hours as shown in the Figure 1.



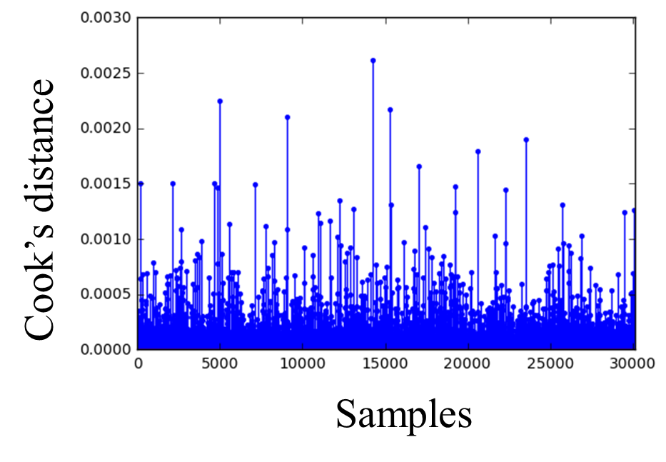
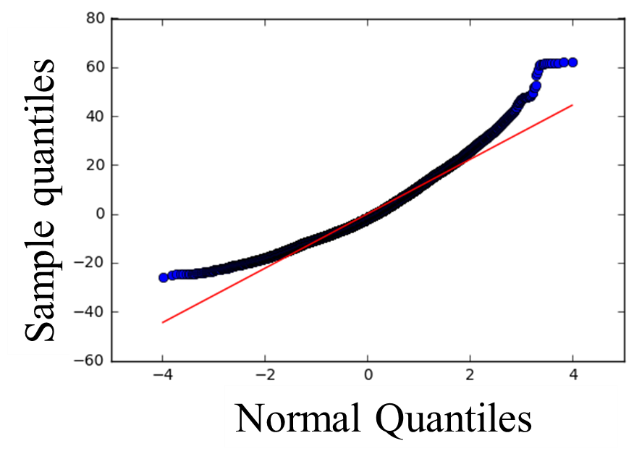
*Figure 1*. A Scatter plot of the Hours per week with respect to the ages. The red line indicate the mean hours per week of individual ages.

1. **What is the difference between adding the “sex” variable in (a) as another predictor compared to only looking at a subset of the data?**

As shown in the Table 2 Column (2), we added the dummy variable of “sex” which is coded male as 1 to the regression model. The coefficients of the sex variable is -0.5563 and the value is statistically significance with the significance level of 0.01. The coefficient indicates that the average age of the female is 0.5563 year older than that of male. The effect of the salary on age is underestimated when only female subset of the data is used.

1. **Are there any outliers in data? Use residual plot and tests to identify regression outliers and influential points.**

As shown in the Figure 2 (left), the Q-Q plot for the residual of the model is not fit to the straight line. Especially, there are many outliers at the right tail of the residual plot. Cook’s distance is utilized to identify the individual outliers if the distance is overhead the threshold which is 4/n where the n is sample size. The number of detected outliers are 1540.

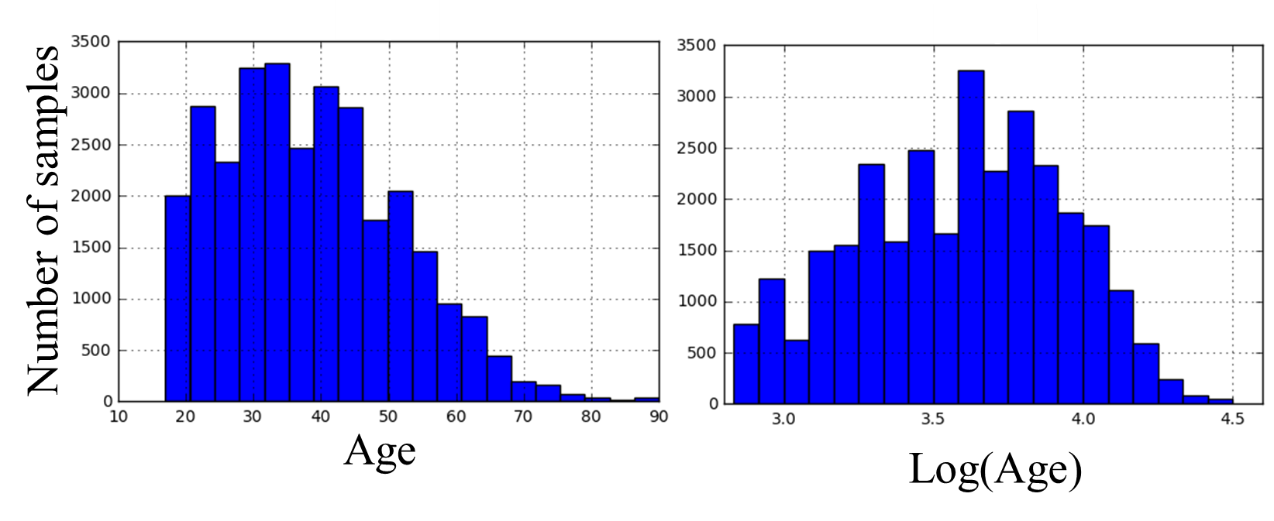


*Figure 2*. A Q-Q plot for the residuals of regression model (left), and the Cook’s distance of the samples (right).

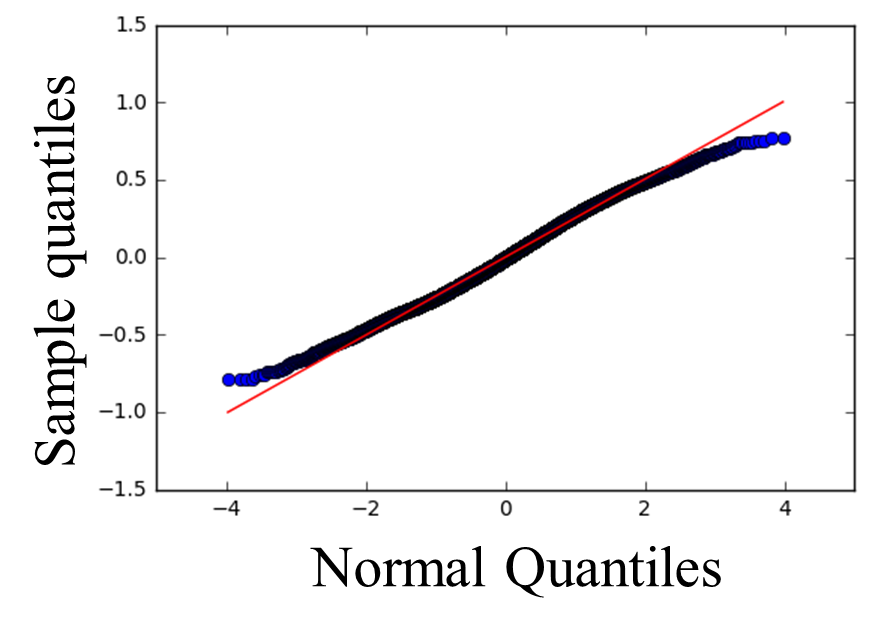
1. **Does the model satisfy the assumptions of regression? Check homoscedasticity, linearity, and normality using plots or tests. Transform the data, as needed, and refit the regression with the necessary modifications.**

The age distribution of samples are also positively skewed as shown in the Figure 3 (left). The age is changed to the log scale to be normally distributed (see Figure 3(right)). To satisfy further assumption of regression, we removed the samples of older than 80 which is irrelevant to the value of hours per week from Figure 1 and the outlier test from the question (e).

The re-fitted regression model is described in the Table 3 Column (3). Every coefficient affects more on the regression on the age because all the magnitude of the t-statistic are increased by the transformation of the original data.



*Figure 3*. Age distribution of samples with respect to the age level (left) and log scaled age (right).



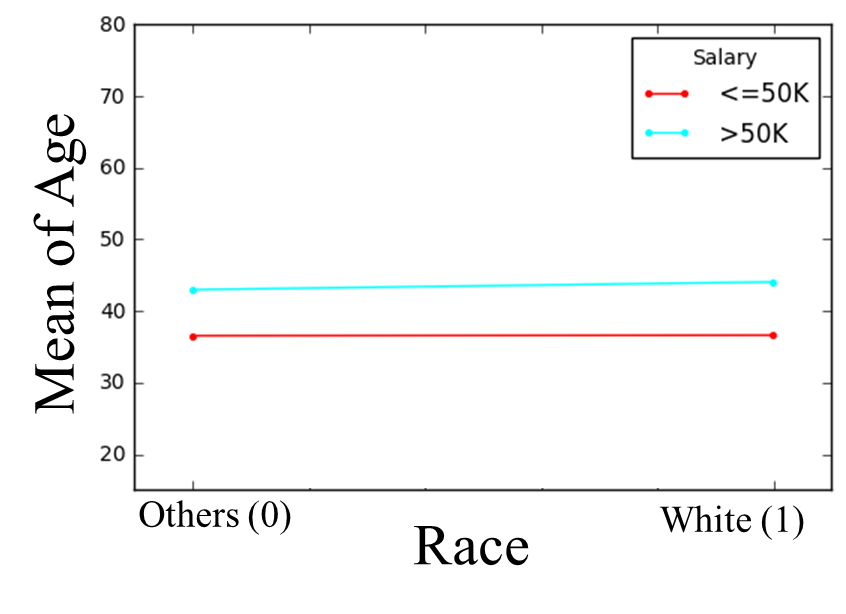
*Figure 4*. Q-Q plot of residual of transformed dataset.

1. **Your instructor is considering whether to add capital gains and losses into the regression model in (e). He believes these two variables will significantly help with all the current variables. Should he or should he not? Assume that the significance level is .05.**

The Table 3 Column (4) indicates that result of the regression analysis adding the capital gain and losses. Adding the capital changes does not affect the other coefficients which indicates capital changes are not necessarily added into the original regression models.

1. **Your instructor believes that the interaction between race and salary may play a crucial role in explaining the variation in age specified by the model in (e). Test if the interaction term is necessary. Assume that the significance level is .01.**

As shown in the Table 3 Column (5), interaction term is added into the regression model used to problem (g). The t-statistic of the interaction term indicates that it is hard to reject null hypothesis because of high p-value of 0.032 compared to the significance level of 0.01. The Figure 5 illustrated that the two lines are parallel, which is there is no interaction effect between race and salary.



*Figure 5*. Interaction plot for Age between race and salary

1. **Check for multicollinearity among the investigated variables.**
2. **Print the correlation table involving age, education, sex (set the reference), HoursPWeek, and salary. Indicate if the correlation is significant.**

Table 4 illustrates the correlation values among the variables. The highest correlation value is 0.632 which is between log scaled age and marital status. However, the correlation does not matter in the regression model since one is dependent variable and the other is regressor. We further investigated variance inflation factor(VIF) as well. A rule of thumb is that if VIF>10 then multicollinearity is high. The VIF value of Hours per week and the education year are scored larger than 10, which is 13.1 and 11.7 respectively. However, we cannot say that there are severe multicollinearity problem in the regression model because the correlation with the other dependent variables are not significantly high.

Table 4

*Correlation table of the variable*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | log(Age) | HoursPWeek | Marital | Sex | Race | EducationYr | Salary | VIF |
| log(Age) | 1.000 | 0.259 | **0.632** | 0.115 | 0.025 | 0.117 | 0.315 |  |
| HoursPWeek |  | 1.000 | 0.253 | 0.243 | 0.062 | 0.159 | 0.239 | 13.1 |
| Marital |  |  | 1.000 | 0.181 | 0.074 | 0.045 | 0.333 | 3.6 |
| Sex |  |  |  | 1.000 | 0.103 | 0.004 | 0.222 | 3.4 |
| Race |  |  |  |  | 1.000 | 0.041 | 0.086 | 6.6 |
| EducationYr |  |  |  |  |  | 1.000 | 0.339 | 11.7 |
| Salary |  |  |  |  |  |  | 1.000 | 1.6 |