**Predictive Analytics Project:**

**What determines whether a person can earn more than $50K per year or not?**

**Executive Summary**

The raw dataset (csv form) is extracted from the 1994 Census Bureau Databased) and obtained from Kaggle (<https://www.kaggle.com/uciml/adult-census-income>), containing relevant variables contributing to predict whether a person’s earning is greater $50K per year. The key variable of interest is a dummy variable, indicating whether a person earns more than $50K a year. The raw dataset has been cleaned in project I into a tidy dataset to be used in this project.

The key variable of interest if a dummy variable, illustrating whether a person’s annual earnings is greater than $50K. The tidy dataset contains a total of 30162 observations, 19 variables among which 12 variables are integers and 7 variables are characters. Results from Project I indicate that income is affected by lots of factors, among which years of education, hours per week, gender and age are main determinants and occupation and work class may also play a role.

This project develops linear regression models to predict whether a person earns more than $50K per year. Training and testing dataset are split randomly into 80% train and 20% test.

**Proposals for 4 Linear Regression Models**

Based on results of Project I, the following four linear regression model is developed:

Model 1

where is the indicator variable regarding whether a person earns more than $50K per year;

is years of education;

is working hours per week;

is dummy variable (=1 if male; =0 otherwise)

is years of age;

Model 1 includes the most relevant variables as is identified in Project I.

Model 2

Model 2 includes most relevant variables as is identified in Project I as well as interactive variables between gender and years of education.

Model 3

Model 3 includes most relevant variables as is identified in Project I, interactive variables between gender and years of education and polynomial terms of age to capture potential nonlinear impact of age.

Model 4

where is working class dummy variable (there are total 7 working classes and the working class of without-pay is excluded to avoid perfect collinearity problem and the rest 6 working class dummies are included);

is occupation dummy variable (there are total 14 occupations and armed forces is excluded to avoid perfect collinearity problem and the rest 13 working class dummies are included).

Model 4 includes as many relevant variables as possible.

**Diagnostic Results**

Regression Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m1 | m2 | m3 | m4 |
| age | 0.0066\*\*\* | 0.0065\*\*\* | 0.0251\*\*\* | 0.0243\*\*\* |
| edu\_yr | 0.0520\*\*\* | 0.0318\*\*\* | 0.0294\*\*\* | 0.0154\*\*\* |
| hr\_week | 0.0046\*\*\* | 0.0047\*\*\* | 0.0034\*\*\* | 0.0029\*\*\* |
| male | 0.1590\*\*\* | -0.1264\*\*\* | -0.1279\*\*\* | -0.0986\*\*\* |
| male.edu\_yr |  | 0.0282\*\*\* | 0.0284\*\*\* | 0.2605\*\*\* |
| age2 |  |  | -0.0002\*\*\* | -0.0002\*\*\* |
| workclass\_Federal-gov |  |  |  | 0.1367 |
| workclass\_Local-gov |  |  |  | 0.0522 |
| orkclass\_Private |  |  |  | 0.0885 |
| workclass\_Self-emp-inc |  |  |  | 0.1933 |
| workclass\_Self-emp-not-inc |  |  |  | 0.0587 |
| workclass\_State-gov |  |  |  | 0.0356 |
| occupation\_Adm-clerical |  |  |  | 0.1042 |
| occupation\_Craft-repair |  |  |  | 0.0973 |
| occupation\_Exec-manageri |  |  |  | 0.0522 |
| occupation\_Farming-fishi |  |  |  | 0.0061 |
| occupation\_Handlers-clea |  |  |  | 0.0450 |
| occupation\_Machine-op-in |  |  |  | 0.0522 |
| occupation\_Other-service |  |  |  | 0.0683 |
| occupation\_Priv-house-se |  |  |  | 0.0815 |
| occupation\_Prof-specialt |  |  |  | 0.2388 |
| occupation\_Protective-se |  |  |  | 0.2195 |
| ccupation\_Sales |  |  |  | 0.1598 |
| occupation\_Tech-support |  |  |  | 0.1827 |
| occupation\_Transport-mov |  |  |  | 0.0762 |
|  |  |  |  |  |
| constant | -0.8236\*\*\* | -0.6180\*\*\* | -0.8970\*\*\* | -0.9362\*\*\* |
| R-squared | 0.2192 | 0.2249 | 0.2357 | 0.2662 |
| Adjusted R-squared | 0.2190 | 0.2247 | 0.2355 | 0.2654 |
| Residual standard error | 0.3832 | 0.3819 | 0.3792 | 0.3717 |
| p-value of F-statistics | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Note: \*\*\* p-value=0.000

Table 1 Regression Results

Coefficients of age, years of education and hours per week are all positive and significant for all models. This is consistent with my expectation because the greater than age/years of education/hours per week, the greater earnings are expected because we can expect more working experiences (greater age), skills (more education) and payment (the more you work, the more you get).

Coefficient of male is positive and significant in model 1: being a model is associated with higher probabilities of earning more than $50K. However, after we add in the interactive terms of male and years of education, the coefficient of male turns into negative for model 2/3/4 but the coefficients of the interactive term are positive. The implications this phenomenon is that it is not always the case that males are more likely to earn more than $50K. Instead, males that more likely to earn greater than $50K must have many years of education (so than the negative impact of male can be compensated).

For model 3 and 4, the coefficients for age is positive but for age square is negative, indicating that the probability that people earn more than 50K increases first with age but would finally decrease if age is too old: the relationship is not linear. This is consistent with my expectation and is the main reason why I establish a model with a polynomial term of age.

For model 4, the coefficients for different working classes and occupation are positive. However, they are not significant even at a 10% level. This is somewhat not consistent with my expectation. I expect at least some working classes/occupations would statistically impact the probability of earning more than 50K.

In terms of goodness of fit, all model’s (adjusted) R-squared are relatively low: about 0.2, but adjusted R-squared (AdjR2) does increase as more explanatory variables are added (AdjR2: m1<m2<m3<m4), indicating that the model 4 has the best goodness-of-fit among the four models.

In terms of model significance, all models are statistically significant.

Based on performance of goodness-of-fit and residual standard error, model 4 has the best performance for the training dataset.

More specifically, the regression equations for model 1 to model 4 are respectively:

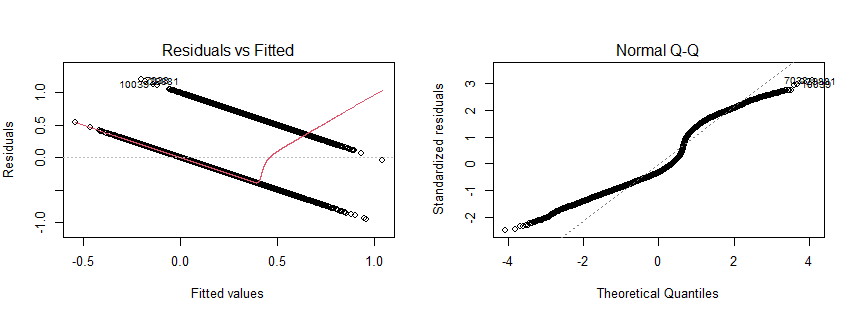
Model 1:

Model 2:

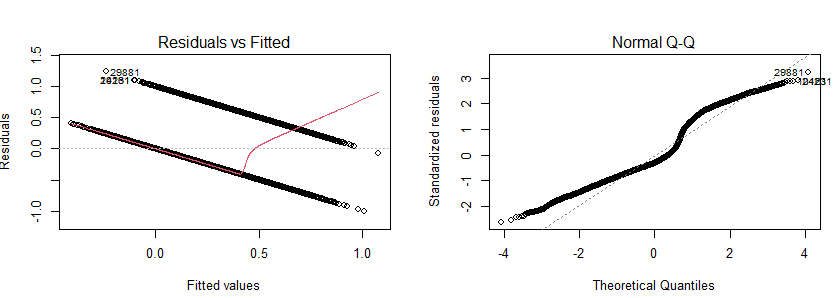
Model 3:

Model 4:

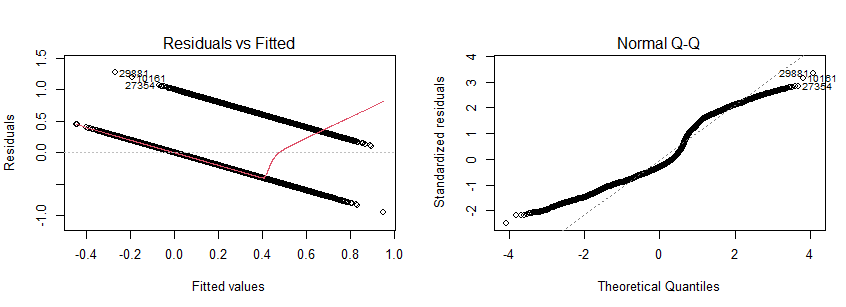
Residual Analysis



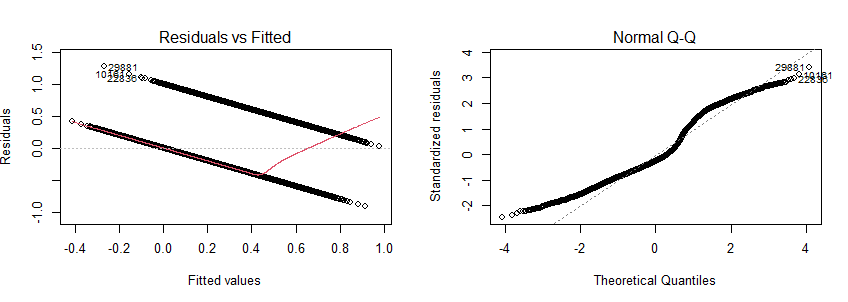
Graph 1 Residual Analysis for Model 1



Graph 2 Residual Analysis for Model 2



Graph 3 Residual Analysis for Model 3



Graph 4 Residual Analysis for Model 4

The normal Q-Q plot in the above residual analysis graph above for 4 models illustrates that the residuals are not normally distributed: instead of forming a closely straight line on the plots, they show an “S” shape. The residuals vs fitted plot, in addition, illustrate that there are non-linear patterns, the residuals clustered into two downward-sloping trends in the plot.

**Predictions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | m1 | m2 | m3 | m4 |
| RMSE | 0.3815 | 0.3801 | 0.3777 | 0.3706 |

Table 2 Prediction Performance

I use root mean square error (RMSE) to measure the out-of-sample performance of my model. Model 4 has the smallest RMSE among the four models. It is also the model (Model 4) that performs the best on the training data (largest (adjusted) R-squared and smallest residual standard error).

**Conclusion and Discussion**

Based on the in-sample performance and out-of-sample performance, I propose that model 4 is the best model among the four multi-linear models. However, the residual analysis indicates that there are potential violations of normality and non-relevance assumption of residuals. Also, our dependent variable is not a continuous variable. Instead, it is a binomial variable with only two values (0 and 1). A multi-linear model may not be good enough to explain variations in the dependent variable (for example, our predicted value are some numbers varying from zero to 1 instead of either 0 or 1). Better non-linear models may be considered in the future (such as logistic regression model).