Final Project: Factors Affecting the Income

Yuwei Zhang

Summary

Income inequality in the United States kept rising in recent years and there are many factors that may have an impact on income, like age, education level, gender, etc. The question naturally arises: **What are the most influential factors in the income of American adults?** That's the goal of this study, which is **inference based**. In addition, I'm also interested in the odds ratio of over $50k annual income **differs by one's age or across native countries. EDA** was conducted to check the association of each predictor variable with the **response variable, whether the annual income exceeds 50k**, and highlighted the preliminary concerns based on the results of the EDA. **Box plots and binned plots** were used to analyze the association between the **binary response variable with each numeric predictor variable**, and **tables** were used to analyze the associations between the **binary response variables with each categorical predictor variable**. Chi-square test is used to compare the deviance of models with and without predictors. VIF is used to check multicollinearity and eliminate redundant variables. The **potential interactions** of the variables are explored to identify differences in data trends for different groups of predictors. **Data modeling and model assessment** are conducted to identify the optimal model to answer key interests. In addition, binned residual plots were used to verify the **assumptions of the final regression model**. Below are important results from the study (more detailed explanation will be provided in Model and Conclusion sections): The outcome of the study shows **age, sex, family relationship, weekly working hour and its square, education level, occupation and native country are all significant predictors** that have an impact on income. People are **less likely to have over 50k income** if age changes from 'young' to **'middle' or 'old'**, but **more likely to have over 50k income** if age changes from 'young' to **'senior'**. It seems **'senior' age are most likely to have a high income**. Furthermore, there are also **interactions of age by relationship and age by weekly working hours**.

Introduction

To analyze the income factors, the relevant data of US Adult Census Income data in 1994 was used, which is obtained from the UCI Machine Learning Repository. The questions I focused on are:

Q1: What are the main factors that have an impact on income?

Q2: Are people more likely to have an annual income of over $50K annual income with larger age?

Q3: Did the overall odds of over 50k annual income differ by native country?

Q4: Are there any other relationships with the odds of over $50K annual income?

The raw data has **32,561 observations** in total. Consistent with the fact that people with high income are in the minority, the number of those who have annual income is > $50k is 7,841, while <=$50k is 24,720. A small amount of data has missing values in both 'type_employer' and 'occupation', I only kept the ones with complete observations for variables I'm interested in.

The response variable is 'income', which is categorized into **binary as '<=50k' and '>50k'**, denoted as 0 and 1 respectively.

At the beginning, there are 15 predictor variables, among which categorical ones are 'type_employer', 'education', 'marital', 'occupation', 'relationship', 'race', 'sex', 'country' and numeric ones are 'age', 'fnlwgt', 'education_num', 'capital_gain', 'capital_loss', 'hr_per_week.'. Some of the predictor variables are telling the same story, thus only the more representative ones were kept. Both 'type_employer' and 'occupation' indicate the job type and **'occupation' was kept** for being more detailed. **Education-related** variables 'education' and 'education_num' are concluded as **5 levels** according to different education stages ('No HS-grad': without high school diploma, 'HS-grad': with high school diploma, 'College': with associate or college degree, 'Bachelors': with Bachelor's degree, 'Masters+': with master or higher degree). 'Marital' and 'relationship'

both indicate the **family status, only the 'relationship' was kept**. Furthermore, the level "Husband" and "Wife" in 'relationship' are merged into 'Married', while level "Not-in-family" and "Other-relative" were merged into 'Unmarried'.

Numeric predictors were also preprocessed. As I mainly focused on the adult income, only the observations with age between 17 and 80 are kept, and the ones **who don't have income are excluded**, that is, 'Without-pay', 'Never-worked' for 'type_employer'. The relationship between age and income is not linear, which is more like an S-curve. Compared to the square of age, **categorizing age** into **'young', 'middle', 'senior' and 'old'** could build a better model. The 'capital_gain' and 'capital_loss' are summed up as differences in the capital, denoted as 'diff'.

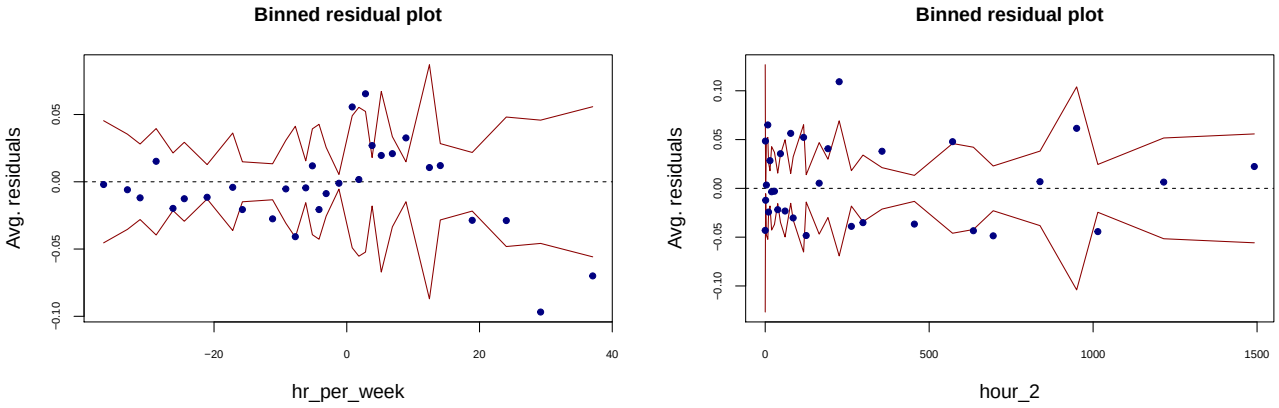
So far, **categorical predictors** are **"age","education","occupation","relationship","race","sex", "country"**. **Numeric predictors** are **"diff","fnlwgt","hr_per_week"**, which are all centered.

DATA

The table of response variable 'income_50k' shows that 24,720 observations have an income less than 50k while 7,841 observations have an income greater than 50k. The negative observations are approximately 24% of all observations, which is still within a reasonable scale.

response variables

The binned plot of 'hr_per_week' against income shows a quadratic trend. Thus transformation is conducted, the binned residual plot is almost random after adding the quadratic term of working hours, demoted as 'hour_2'



From the boxplot of capital difference 'diff' against income, the quantiles are all 0 with a few outliers, which is reasonable as most observations have 0 capital_gain nor have capital_loss. There's one **outlier who has an extremely large capital gain ($100,000)** compared to all other observations and was **dropped**.

The boxplot of census generated variable 'fnlwgt' against income showed almost no difference, thus 'fnlwgt' was eliminated.

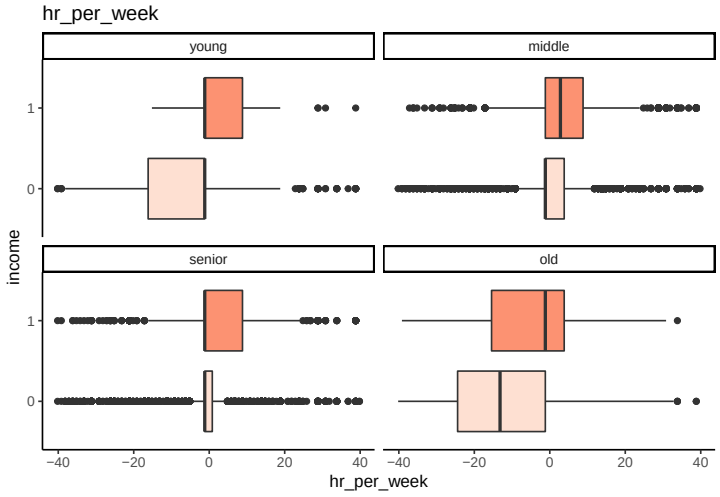
As for categorical predictors, table and chi-test are used to estimate the relationship. It turns out that **all categorical variables** have a corresponding p-value much smaller than .05, indicating **some correlation with the response variable**. Some **native countries** with few observations could borrow information from other native countries, but 'Holand-Netherlands' was dropped for only one observation.

To check the multicollinearity, a basic logistic regression model was conducted with all predictors. The VIF of white race and black race were both larger than 10, thus the race is eliminated from the predictors.

The 'country' indicating **native country** brings some **hierarchical features** and wass added into the model as a **random intercept**. Thus numeric predators 'hr_per_week', 'diff', and all categorical variables except for 'country' were included in our model.

interactions

Box plots and tables are used to identify the **potential interactions** between two predictors versus ratio. Due to the length limitation of the report, only the interesting and important interaction findings are highlighted. There are 5 potential interactions: **age by relationship, age by education, age by occupation, age by hr_per_week, sex by education**. To determine the significance of those interactions, **anova tests** were later conducted during model fitting.



At the end of EDA, there are 29,347 remaining observations after dropping out the outliers and some predictors.

Model

Based on the EDA results, the preliminary model was constructed with predictor variables: **'hr_per_week', 'hour_2', 'diff', 'age', 'education', 'relationship', 'sex' and 'occupation'**. Since the **response variable 'income_50k' is categorical ('<=50k' and '>50k')**, it's suitable to use the **logistic regression** to fit the income levels. As discussed above, **'country' brings some hierarchical features**, thus a **multilevel model** was adopted to better explain the data with the random intercept of the native country.

The summary of the preliminary model showed that predictor variables 'hr_per_week', 'hour_2', 'diff', 'age_fac', 'education', 'relationship', 'sex' and 'occupation' all have significant relationship with response variable. The **standard deviation of random intercept 'country' was 0.37**, indicating **some variation explained by native countries**. The dotplot proved that the overall odds of income over 50k differs by native countries.

Then, the **potential interactions** from EDA were checked by **anova test**. There were 2 interactions having p-value smaller than .05 and were added into the final model. The remaining 2 significant interactions were both related to age, they were **age by relationship** and **age by weekly working hours**.

final model

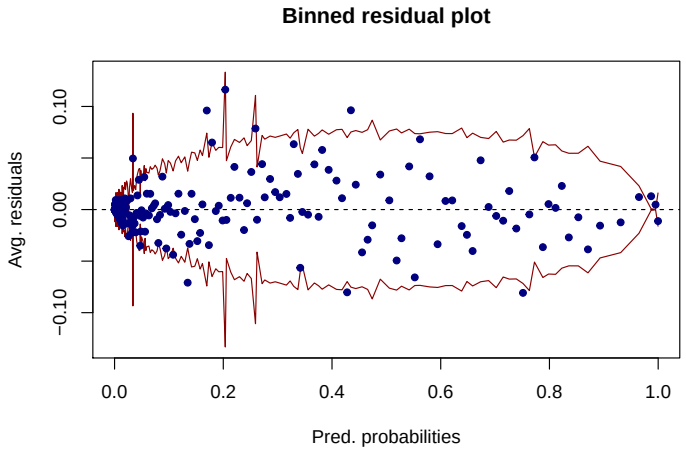
Based on above work, we finalized our model as:

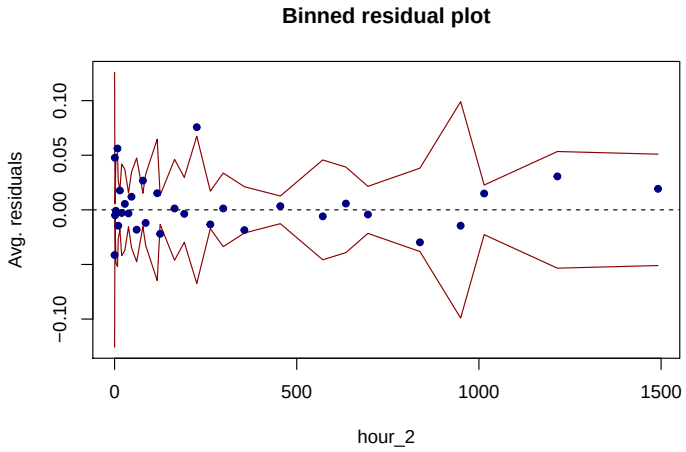
$$log(\frac{\pi_{ij}}{1-\pi_{ij}}) = (\beta_0 + \gamma_{0j}) + \beta_1 age\_fac_{ij} + \beta_2 sex_{ij} + \beta_3 relationship_{ij} + \beta_4 hr\_per\_week_{ij} + \beta_5 hour\_2_{ij} + \beta_6 diff_{ij}$$
$$+\beta_7 education_{ij} + \beta_8 occupation_{ij} + \beta_9 age\_fac : relationship_{ij} + \beta_{10} age\_fac : hr\_per\_week_{ij} + \epsilon_{ij}$$
$$where \pi_{ij} = income over 50k, \epsilon_{ij} \sim N(0, \sigma^2), \gamma_{0j} \sim N(0, \tau_0^2)$$

The predictors for final model are 'hr_per_week', 'hour_2', 'diff', 'age_fac', 'education', 'relationship', 'sex' and 'occupation'. With 2 interactions age by relationship and age by weekly working hours, plus random intercept of native countries.

model validation

Binned residual plots were plotted to validate our model. The points in binned residual plots are almost randomly distributed, thus the random assumption is well suited. Few points are out of the 95% bands and the model is well fitted.





To check the model's performance on data,we plot confusion matrix and ROC curve. The sensitivity and specificity of the model are 0.58 and 0.93 respectively, with a high AUC of 0.895, indicating the model is well fitted.

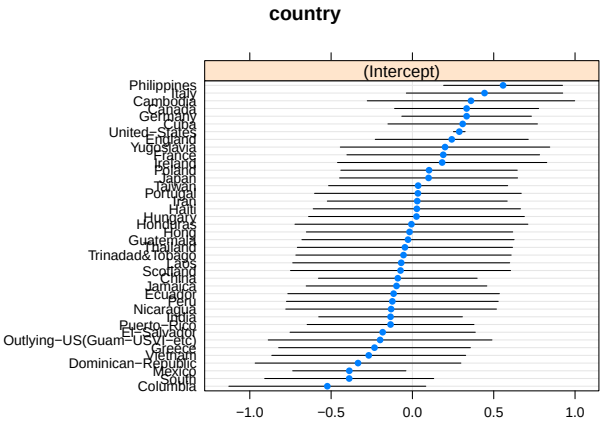
Conclusion

Predictors	Odds ratio	Predictors	Odds ratio
intercept	0.73	sex	1.20
age[middle]	2.28	relationshipOwn-child	0.00
age[senior]	0.33	relationshipUnmarried	0.12
age[old]	0.98	hr_per_week	1.04
age[mid]:relationshipOwn-child	0.00	diff	1.00
age[sen]:relationshipOwn-child	0.00	education[College]	1.00
age[old]:relationshipOwn-child	0.70	education[HS-grad]	0.49
age[mid]:relationshipUnmarried	2.19	education[Masters+]	0.33
age[sen]:relationshipUnmarried	1.26	education[No HS-grad]	1.73
age[old]:relationshipUnmarried	1.10	occupation[Exec-managerial]	0.15

Predictors	Odds ratio	Predictors	Odds ratio
age[mid]:hr_per_week	0.96	occupation[Prof-specialty]	2.05
age[sen]:hr_per_week	1.02	occupation[Tech-support]	1.81
age[old]:hr_per_week	0.99	occupation[Farming-fishing]	0.31

According to the summary of our final model, all individual predictors have p-value smaller than .05 and have significant relationship with response variable 'income_50k'.**The individual predictors in final model are all main factors that have an impact on income (Q1)**. The baseline intercept of the final model shows that a young married female from India with bachelor's degree who works for 0 hour a week as a clerk and has 0 capital income has the odds of over 50k income be 0.72.

The odds of over 50k income would **increase by 128% (confidence interval 74%-198%)if the age changes from 'young' to 'middle'**, with all other variables constant. But if the age changes from **'young' to 'senior' or 'old', the odds of over 50k income would decrease by 67% (confidence interval 59%-73%) and 2% (confidence interval -9%-12%)**. Thus the **odds of over 50k income would first increase and then decrease with larger age (Q2)**. **On top of the effect of age, the interactions related to age also have an impact on income.** According to the **interaction of age by relationship**, the odds of over 50k income differ with different education levels. If a married one who has 'middle' age changes to have a child, the odds of over 50k income would decrease by 100%, while changes to unmarried would increase the odds by 119%. For **interaction of age by weekly working hours**, if the baseline has already changed into 'middle' age, the one unit increase in weekly working hours would decrease the odds of over 50k income by 4%, with all other variables constant. However, if the age is 'senior', the one unit increase in working hours would lead to an increase of 2% in odds.



The **random intercept helps to explain variations across native countries**. The cross-country variation attributed to the random intercept of 'country' is 0.377. The dotplot shows that **different native countries have different intercepts (Q3)**. Among all countries, only the confidence interval of 'mexico' and 'united states' did't include the 0 intercept, which were different from other countries. This makes sense because the income census was based on the United States and those whose's native countries are the United States or Mexico have fewer policy restrictions concerning the job market.

limitations

There are also some potential limitations of the study. Firstly, the income census is not the latest years, there might be some change in the job market and policy. Besides, though it's within a reasonable scale, the data is **not quite balanced** for positive and negative observations. The number of positive observations are about 3 times of the negative.