Predictive Factors for Income Level: A Multiple Logistic Regression Analysis

Dallin Johnson

Department of Biostatistics
University of Kansas
May 10, 2023

Abstract

Inflation and rising costs in the United States have inspired the main purpose of this study: which factors or combination of factors are the most important in determining income? The data that was analyzed for this study contains multiple variables that helped describe the individuals within the dataset. Income was one of the variables in this dataset which grouped every individual into two categories: those who made more than $50,000 and those who made less than or equal to $50,000 that year. One of the main reasons that a statistical analysis was conducted on this particular set of data is because I believed that age and education were the predominant factors in determining whether someone will make more than $50,000 a year in the United States. I predicted that the older the individual, combined with the more schooling that this particular person had, would have a positive correlation with the probability that said individual would earn a salary larger than $50,000.

This original dataset was initially extracted by Barry Becker in 1994 from a "Census database" (Dua and Graff 2019). This data was then published in the University of California, Irvine Machine Learning Repository. Since the main question of the analysis included a binary dependent variable, it was decided that a multiple logistic regression model would be the best choice to predict income. The final model included several variables such as age, workclass, education, race, sex, and hours per week (the number of hours an individual worked per week) in order to predict the dependent variable of income. It was found that the variables race, workclass, and hours per week in the final model were the variables that had the strongest correlation with the dependent variable. Another crucial finding was that the absolute value of the z-score for age and some of the categories that belonged to the education variable were very large, indicating that there was a strong correlation between them and predicting an individual's income.

Introduction

The reason this study was conducted stemmed from the economic hardships that plague many Americans today. According to MIT's Living Wage Calculator, the living wage for an adult in the US is roughly $33,000 (MIT 2023). While acknowledging that $50,000 is a considerably higher salary than $33,000, this analysis deemed it appropriate to utilize the former as a benchmark for analyzing whether individuals were earning in excess of that amount. As indicated earlier in this analysis, my initial assumption was that an individual's age and education would have the strongest correlation with their income. I proposed that age and education would exhibit a greater correlation with an individual's income, as it is widely observed that as an individual progresses in age, they are likely to receive higher pay. Additionally, I have noticed that many high-paying positions require graduate and undergraduate degrees.

While it is assumed that the source of the census database was the U.S. Census Bureau, extensive digging failed to prove what the original source of the data was. The primary limitation of the data involves the income variable, with responses indicating whether a person earns more than $50,000 annually or whether they earn less than or equal to $50,000 annually. There is a wide range of salaries throughout the US, and grouping individuals into two categories does not paint the best picture of the diversity in salaries. Other limitations found in this dataset pertain to the variables of occupation and Race. The occupation variable contains many different job types, some of which are poorly defined. Race is another categorical variable that is not comprehensive, as the dataset contains only five race options, none of which include Hispanic/Latino, despite

their significant representation in the United States population. Despite the limitations, the abundance of observations available for analysis, with a total of 32,560 observations included, helped make this dataset great for analysis.

Methods

In order to gain a comprehensive understanding of the dataset analyzed, it was necessary to find out more about the variables. Firstly, the dependent variable was income, which is a categorical variable comprising of two categories: one indicating whether an individual earned over $50,000, and the other indicating if they earned less than or equal to $50,000. The independent variables that were used for this analysis include age, workclass, education, marital-status, occupation, relationship, race, sex, hours per week, and native country. All of these independent variables are categorical except for the age and the hours per week variables which are continuous. To provide a more comprehensive overview of the categories within the categorical variables, a table has been included below that displays the specific categories within each variable:

Table 1: Independent Categorical Variables

Variable	Categories
Workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Marital-Status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
Relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black

Sex	Female, Male
Native Country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands

All observations with null values were eliminated from the dataset. This decision was made based on the fact that the null values represented less than 10% of the total observations and were randomly distributed. A filter was also applied to include only individuals whose native country was classified under the category of the United States. While the data may provide insights for individuals not native to the United States, it was uncertain if the dataset included census data from other countries and this data was filtered to avoid any mixing of data or census from other countries with the data found within the US. Filtering the data set to include only individuals from the United States resulted in a reduction of less than 10% of the total number of observations. The income variable was subsequently converted to binary numbers, with one denoting adult income over $50,000 and zero denoting adult income less than or equal to $50,000.

After completing the data cleansing process, descriptive statistics were then calculated from the variables within the dataset. Firstly, the mean, median, and standard deviation of all the continuous variables used for the logistic regression model were calculated. The results of these calculations are shown in the table below:

Table 2: Descriptive Statistics

Variable	**Mean**	**Median**	**Standard Deviation**
Age	38.5	37	13.2
Hours per Week	41.0	40	12.0

In order to further visualize the continuous variables, these histograms were created as well:

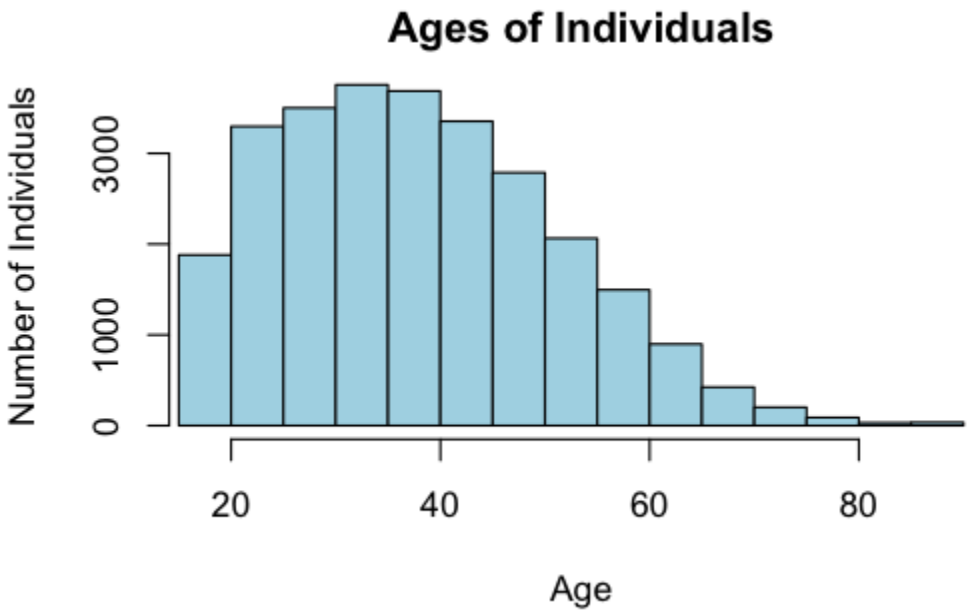

Figure 1: Distribution of Age

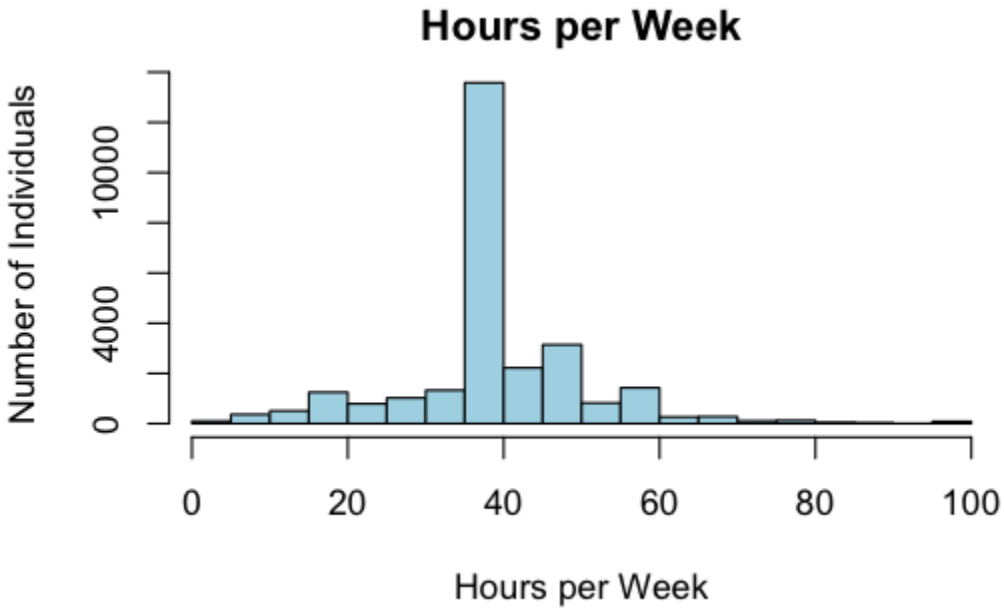

Figure 2: Distribution of Hours Worked per Week

The proportions of each category within each categorical variable were also calculated, however, it was decided that this information would not be included in the following paper due to its length. Rest assured, these proportions were taken into account with the analysis of the data.

Concerning multicollinearity, a correlation analysis was conducted using income as the dependent variable and the following independent variables: age, workplace, education, marital-status, occupation, relationship, race, sex, and hours per week. A generalized linear model of a binomial family was constructed using these variables. Then, a variance inflation factor (VIF) analysis was conducted to detect multicollinearity in the multiple regression model. Several categorical variables had VIF values exceeding the conventional threshold of 2.5. Specifically, relationship and marital status had extremely high VIF values, with relationship having a VIF value of approximately 133.6 and marital status having a VIF value of 63.5. This came as no surprise since it makes sense that Relationship and Marital-Status are closely related variables. Also, occupation and sex both had VIF values that were close to 2.7, which was larger than the conventional threshold. Considering these results, it was determined that sex would remain in the dataset since its VIF value was so close to the 2.5 thresholds. However, occupation was removed due to its potential to contribute to overfitting and decrease the accuracy of future predictions. The occupation variable added 13 degrees of freedom to the model and had categories that were difficult to interpret.

A backward stepwise selection procedure was then performed on the remaining variables, and the Akaike Information Criterion (AIC) was used to assess the quality of these remaining variables. Education was found to have the highest AIC value, meaning it was the least reliable at providing the goodness of fit compared to the other variables. However, education was a key factor in the study and my hypothesis, so it was decided to keep education in the final model. The final multiple logistic regression model comprised income as the dependent binomial categorical variable, while age, workclass, education, race, sex, and hours per week serve as independent variables. Some of the categories under education, age, sex, and hours per week produced some of the largest absolute values for their z-scores. This suggests that there was a stronger relationship between these independent variables and categories when compared to the other independent variables and categories.

A chi-squared test was conducted on the final multiple logistic regression model. All of the independent variables used in the model had a p-value below the 0.05 threshold, indicating statistical significance. Race had the highest p-value of 0.0188. This provides sufficient evidence to reject the null hypothesis for each of the variables, meaning there is a correlation between these independent variables and the dependent variable.

The coefficients were then extracted from the final multiple logistic regression model, which was used to produce a logarithmic equation that predicts the likelihood of an individual's income is greater than $50,000, or is equal to or less than $50,000:

logit(p) = -6.9636 + 0.0473*Age - 0.4722*WorkclassLocal-gov - 0.3956*WorkclassPrivate + 0.1062*WorkclassSelf-emp-inc - 0.8767*WorkclassSelf-emp-not-inc - 0.6232*WorkclassState-gov - 12.8372*WorkclassWithout-pay + 0.0688*Education11th + 0.4713*Education12th - 1.4908*Education1st-4th - 1.2778*Education5th-6th - 0.6403*Education7th-8th - 0.3323*Education9th + 1.7354*EducationAssoc-acdm + 1.6802*EducationAssoc-voc + 2.4034*EducationBachelors + 3.4848*EducationDoctorate + 0.9997*EducationHS-grad + 2.851*EducationMasters - 10.6401*EducationPreschool + 3.4665*EducationProf-school + 1.4095*EducationSome-college + 0.4586*RaceAsian-Pac-Islander + 0.1199*RaceBlack - 0.012*RaceOther + 0.538*RaceWhite + 1.1915*SexMale + 0.0333*Hours-per-Week

A multiple logistic regression plot was then constructed using the predicted values derived from the final model, juxtaposed with the actual values present in the dataset. The predicted values were plotted along the y-axis, while the actual value, were plotted along the x-axis:

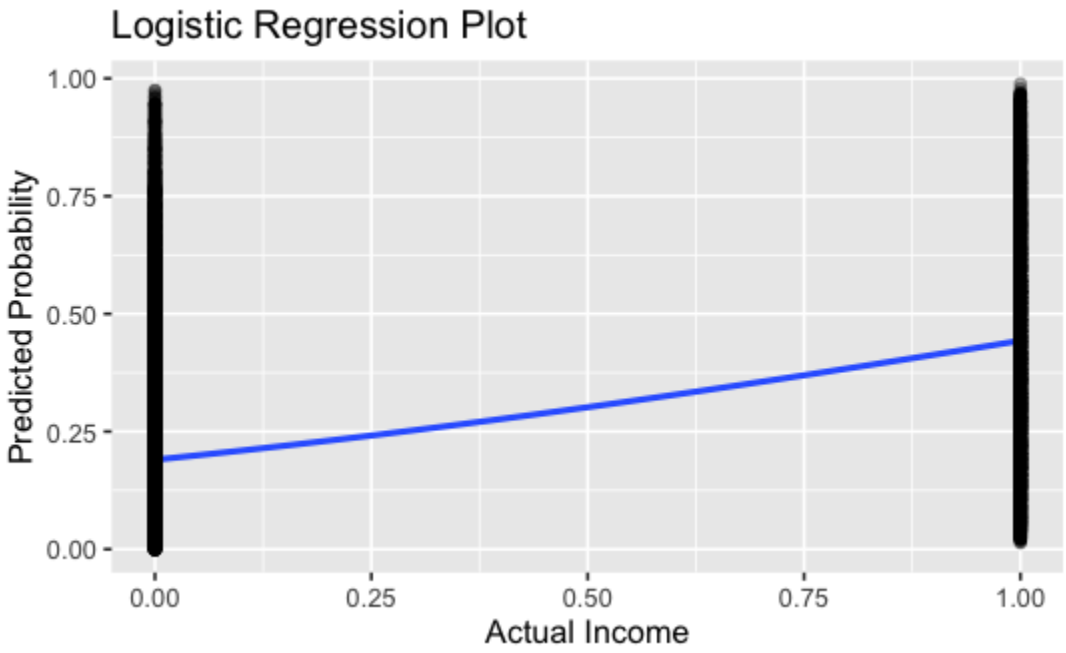

Figure 3: Final Multiple Logistic Regression Plot

A confusion matrix to evaluate the accuracy of the model was then created. The confusion matrix is presented below:

Table 3: Accuracy of Final Model

Prediction	**Actual Value**	
	Less than or Equal to 50K	**Greater than 50K**
Less than or Equal to 50K	19,116	4,192
Greater than 50K	1,392	2,803

The presented matrix displays the predictions generated by the final multiple logistic regression model, organized as rows, against the actual values of the independent variable, arranged as columns. There were 19,116 true negatives, 4,192 false negatives, 1,392 false positives, and 2,803 true positives. The model's accuracy was determined by calculating the percentage of correct predictions, which rounded to the nearest tenth, was found to be 79.7%.

Another confusion Matrix was then created for a model that included all of the independent variables from the beginning of the analysis (basically including all of the variables that were deemed unfit for the final model) which is shown here below:

Table 4: Accuracy of Full Model

Prediction	**Actual Value**	
	Less than or Equal to 50K	**Greater than 50K**
Less than or Equal to 50K	19,020	3,878
Greater than 50K	1,488	3,117

The complete model resulted in an accuracy of 80.5% rounded to the nearest tenth. While it is logical that this model would be more accurate due to the inclusion of more variables, it was decided that there is high confidence in the efficacy of the final model. This simpler model avoids the issue of overfitting the data, which could result in a model that does not generalize well to new data.

This last confusion matrix was similar to the one generated from the final logistic regression model, with the exception that the education variable was excluded. This was due to the fact that the education variable had the highest AIC value. Thus, it was deemed necessary to see how removing education would affect the accuracy of the model:

Table 5: Accuracy of Final Model Excluding Education

Prediction	**Actual Value**	
	Less than or Equal to 50K	**Greater than 50K**
Less than or Equal to 50K	19,308	5,606
Greater than 50K	1,200	1,389

After conducting an analysis of a model without the education variable, it was found that the accuracy of this model was 75.3% rounded to the nearest tenth. It was determined that the final logistic regression model that included the education variable had a higher accuracy rate compared to this model. Although it is unclear if the inclusion of the education variable could lead to overfitting, this result suggests that it is an important variable to consider when predicting whether an individual's salary exceeds $50,000. Further investigation would probably be needed to better understand the correlation between education and model accuracy.

Discussion

Given that the dependent variable was categorical and dichotomous, a multiple logistic regression model appeared to be the best option in order to see the relationship between the independent and dependent variables. Looking at the independent variables, it was observed that the average age of adults was roughly 38.5 years old, and the average hours worked per week was approximately 41.0 hours. The histogram of the age variable demonstrated that this variable was right-skewed, while the histogram of the hours per week variable demonstrated that most individuals work 40 hours per week, which is standard in the US. The decision was made to not standardize these variables in the model, as it is unnecessary in multiple logistic regression and would complicate the final equation.

The categorical variables showed that the majority of individuals in the data set worked for a private company, were white, male, and had an income of $50,000 or less. It is worth noting that these proportions were filtered to individuals residing in the United States. To ensure the accuracy of the income data, for future studied, it would be recommended that data be extracted directly from the US Census Bureau. The limitations of the race variable categories led to the conclusion that a more accurate depiction of the US population would require additional categories. The occupation variable was somewhat confusing and a more precise depiction of occupation that encompasses all US occupations would be recommended for future study.

The backward stepwise selection was performed on the final variables of the multiple logistic regression model, which provided essential insights into the goodness of fit of the variables. Surprisingly, education did not fit the model as well as other variables, as evidenced by its AIC values. Further analysis revealed that while some categories under the education variable had high absolute z-scores, others, particularly those where the education level was less than college, were not as accurate in predicting the dependent variable.

The final model accurately predicted the dependent variable approximately 79.7% of the time. This is not as accurate as hoped for, however, future studies with more datasets could provide more accurate models to predict income.

Conclusion

In summary, this report details a statistical analysis conducted to identify the variables or combination of variables that are most important in determining whether an individual in the US makes a salary greater than $50,000. The analysis focused on the hypothesis that age and education were strong predictors of income. Education was deemed the least fit for the model when compared to the other independent variables included in the final model. The analysis revealed that race, work, and hours were the variables that fit the model the best, but age and education also showed a high correlation with income.

The report provides a detailed description of the methods used in the analysis, including data cleaning, descriptive statistics, correlation analysis, and backward stepwise selection to select the variables to be included in the final model. The report also highlights the limitations of the dataset and provides recommendations for future studies, such as using more accurate income data from the US Census Bureau and improving variable selection.

Overall, the report demonstrates the effectiveness of the methods used in identifying the variables most strongly associated with income. The findings of this study provide valuable insights that could prove to be useful for future studies.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

"Living Wage Calculator." Massachusetts Institute of Technology, 2023, https://livingwage.mit.edu/.

Appendices

```
#Download Dataset
adult.data <- read_csv("~/Downloads/adult.data")
## Rows: 32560 Columns: 15
## ── Column specification ──────────────────────────────────────────────────────
## Delimiter: ","
## chr (9): State-gov, Bachelors, Never-married, Adm-clerical, Not-in-family, W...
## dbl (6): 39, 77516, 13, 2174, 0, 40
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
#View(adult.data)
#Clean Data and filter for only the US
colnames(adult.data) <- c("Age", "Workclass", "fnlwgt", "Education", "Education-Num", "Marital-Status", "Occupation", "Relationship", "Race", "Sex", "Capital-Gain", "Capital-Loss", "Hours-per-Week", "Native-Country", "Income")
#Removing random NULL values since the null values are randomly distributed throughout the dataset.
adult.data[adult.data == "?"] <- NA
adult.data <- na.omit(adult.data)
#Filter for the United States
united.states <- filter(adult.data, `Native-Country` == "United-States")
#View(united.states)

#Removing variables that will not be used in the analysis
#I'm removing native-country since we are only analyzing people with "US" as their native country
adult.data <- subset(united.states, select = c("Age", "Workclass", "Education", "Marital-Status", "Occupation", "Relationship", "Race", "Sex", "Hours-per-Week", "Income"))
#Converting "Income" variable into a binary variable. If income >50K than = 1, if income <=50K than = 0.
adult.data$Income <- ifelse(adult.data$Income == ">50K", 1, adult.data$Income)
adult.data$Income <- ifelse(adult.data$Income == "<=50K", 0, adult.data$Income)
#View(adult.data)
#Descriptive Statistics
#Mean of continuous variables
mean(adult.data$Age)
## [1] 38.50427
mean(adult.data$`Hours-per-Week`)
## [1] 40.97102
#Median of continuous variables
median(adult.data$Age)
## [1] 37
median(adult.data$`Hours-per-Week`)
## [1] 40
#Standard Deviation
sd(adult.data$Age)
## [1] 13.1846
sd(adult.data$`Hours-per-Week`)
## [1] 12.04134
#Proportions of Categorical variables
work.prop <- prop.table(table(adult.data$Workclass))
education.prop <- prop.table(table(adult.data$Education))
marital.prop <- prop.table(table(adult.data$`Marital-Status`))
occupation.prop <- prop.table(table(adult.data$Occupation))
relationship.prop <- prop.table(table(adult.data$Relationship))
race.prop <- prop.table(table(adult.data$Race))
sex.prop <- prop.table(table(adult.data$Sex))
income.prop <- prop.table(table(adult.data$Income))
#Descriptive Statistic Visuals
#Histograms of Continuous Variables
hist(adult.data$Age, main = "Ages of Individuals", xlab = "Age", ylab = "Number of Individuals", col = "lightblue", xlim = c(min(adult.data$Age), max(adult.data$Age)))

hist(adult.data$`Hours-per-Week`, main = "Hours per Week", xlab = "Hours per Week", ylab = "Number of
```

```
Individuals", col = "lightblue", xlim =
c(min(adult.data$`Hours-per-Week`),
max(adult.data$`Hours-per-Week`)))

#Proportions of Categorical Data
work.prop
##
##     Federal-gov       Local-gov         Private
Self-emp-inc
##    0.0322146675    0.0711195142
0.7321019525    0.0360324328
## Self-emp-not-inc       State-gov     Without-pay
##    0.0840999164    0.0439588409
0.0004726757
education.prop
##
##        10th        11th        12th     1st-4th
5th-6th     7th-8th
##  0.027342472  0.034796204  0.012035051
0.001418027  0.002836054  0.015889176
##         9th   Assoc-acdm   Assoc-voc   Bachelors
Doctorate     HS-grad
##  0.012725884  0.034141730  0.044831473
0.167872596  0.011416936  0.334836200
##     Masters   Preschool  Prof-school Some-college
##  0.053957750  0.000545395  0.017743519
0.227611533
marital.prop
##
##            Divorced     Married-AF-spouse
Married-civ-spouse
##         0.1451841617         0.0007635531
0.4653310548
## Married-spouse-absent         Never-married
Separated
##         0.0084718031         0.3226920700
0.0299967276
##             Widowed
##         0.0275606297
occupation.prop
##
##     Adm-clerical     Armed-Forces     Craft-repair
Exec-managerial
##       0.125368142       0.000327237
0.133985383       0.135803367
##   Farming-fishing Handlers-cleaners
Machine-op-inspct     Other-service
##       0.031960150       0.043231647
0.061338763       0.100970803
##   Priv-house-serv    Prof-specialty   Protective-serv
Sales
##       0.003272370       0.134276261
0.022033960       0.122313929
##      Tech-support  Transport-moving
##       0.030905719       0.054212268
relationship.prop
##
##        Husband  Not-in-family Other-relative
Own-child      Unmarried
##     0.41504563     0.25920809     0.02341563
0.15220158     0.10464313
##           Wife
##     0.04548595
race.prop
##
## Amer-Indian-Eskimo Asian-Pac-Islander
Black              Other
##       0.009853471       0.009926190
0.095589572       0.004108643
##              White
##       0.880522125
sex.prop
##
##    Female      Male
## 0.3247646 0.6752354
income.prop
##
##         0         1
## 0.7456641 0.2543359
#Model Creation and Correlation Analysis
adult.data$Income <- as.numeric(adult.data$Income)
log.model <-
glm(Income~Age+Workclass+Education+`Marital-St
atus`+Occupation+Relationship+Race+Sex+`Hours-p
er-Week`, data = adult.data, family = "binomial")
vif(log.model)
##                     GVIF Df GVIF^(1/(2*Df))
## Age                1.240927  1        1.113969
## Workclass          1.592267  6        1.039524
## Education          1.876473 15        1.021201
## `Marital-Status`  63.541843  6        1.413367
## Occupation         2.779873 13        1.040107
## Relationship     133.628992  5        1.631511
## Race               1.057292  4        1.006988
## Sex                2.732551  1        1.653043
## `Hours-per-Week`   1.132581  1        1.064228
#Took out the variables that were too closely
correlated to other variables in the dataset.
adj.log.model <-
glm(Income~Age+Workclass+Education+Occupation
+Race+Sex+`Hours-per-Week`, data = adult.data,
family = "binomial")
vif(adj.log.model)
##                    GVIF Df GVIF^(1/(2*Df))
## Age                1.106537  1        1.051921
## Workclass          1.569202  6        1.038261
## Education          1.799333 15        1.019773
## Occupation         2.618241 13        1.037713
## Race               1.039638  4        1.004871
## Sex                1.169839  1        1.081591
```

```
## `Hours-per-Week` 1.094229  1        1.046054
#Closer look at the correlation between Marital-status and relationship
relationship.data  <- adult.data[,4:6]
#View(relationship.data)
#It makes sense that the Marital-Status variable would be closely related/correlated to the Relationship variable.
#Prevented overfitting the model.
new.model.one <- glm(Income~Age+Workclass+Education+Race+Sex+`Hours-per-Week`, data = adult.data, family = "binomial")
#removing occupation in order to prevent over fitting since it accounts for so many variables. and it has a VIF over 2.5!
#summary(new.model.one)
# assume your data frame is called 'my_data'
# and your response variable is called 'outcome'

# perform forward stepwise selection
model_forward <- glm(Income ~ 1, data = adult.data, family = binomial())
model_forward <- step(model_forward, direction = "forward")
## Start:  AIC=31193.08
## Income ~ 1
# perform backward stepwise selection
model_backward <- glm(Income~Age+Workclass+Education+Race+Sex+`Hours-per-Week`, data = adult.data, family = binomial())
model_backward <- step(model_backward, direction = "backward")
## Start:  AIC=24169.08
## Income ~ Age + Workclass + Education + Race + Sex + `Hours-per-Week`
##
##                 Df Deviance   AIC
## <none>               24111 24169
## - Race           4    24162 24212
## - Workclass      6    24283 24329
## - `Hours-per-Week` 1    24654 24710
## - Sex            1    25054 25110
## - Age            1    25379 25435
## - Education     15    26845 26873
log.model <- glm(Income~Age+Workclass+Education+`Marital-Status`+Occupation+Relationship+Race+Sex+`Hours-per-Week`, data = adult.data, family = "binomial")
#summary(log.model)
#drop1(log.model, test = "Chisq")
#Final Model
final.model <- glm(Income~Age+Workclass+Education+Race+Sex+`Hours-per-Week`, data = adult.data, family = "binomial")
summary(final.model)
##
## Call:
## glm(formula = Income ~ Age + Workclass + Education + Race + Sex +
##     `Hours-per-Week`, family = "binomial", data = adult.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7234  -0.6785  -0.4184   0.4448   2.9322
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -6.963562   0.275651 -25.262  < 2e-16 ***
## Age                         0.047292   0.001364  34.664  < 2e-16 ***
## WorkclassLocal-gov         -0.472168   0.097640  -4.836 1.33e-06 ***
## WorkclassPrivate           -0.395603   0.081954  -4.827 1.39e-06 ***
## WorkclassSelf-emp-inc       0.106218   0.109079   0.974  0.33017
## WorkclassSelf-emp-not-inc  -0.876656   0.096298  -9.104  < 2e-16 ***
## WorkclassState-gov         -0.623239   0.110898  -5.620 1.91e-08 ***
## WorkclassWithout-pay      -12.837203 128.618757  -0.100  0.92050
## Education11th               0.068793   0.204616   0.336  0.73672
## Education12th               0.471287   0.259891   1.813  0.06977 .
## Education1st-4th           -1.490809   1.029693  -1.448  0.14767
## Education5th-6th           -1.277788   0.612306  -2.087  0.03690 *
## Education7th-8th           -0.640259   0.245417  -2.609  0.00908 **
## Education9th               -0.332308   0.269515  -1.233  0.21758
## EducationAssoc-acdm         1.735436   0.165910  10.460  < 2e-16 ***
## EducationAssoc-voc          1.680234   0.160488  10.470  < 2e-16 ***
## EducationBachelors          2.403429   0.148671  16.166  < 2e-16 ***
## EducationDoctorate          3.484776   0.204159  17.069  < 2e-16 ***
## EducationHS-grad            0.999680   0.147434   6.781 1.20e-11 ***
```

```
## EducationMasters            2.851017   0.156353
18.234  < 2e-16 ***
## EducationPreschool        -10.640078 123.406426
-0.086  0.93129
## EducationProf-school        3.466453   0.183850
18.855  < 2e-16 ***
## EducationSome-college       1.409487   0.148677
9.480  < 2e-16 ***
## RaceAsian-Pac-Islander      0.458556   0.256220
1.790  0.07350 .
## RaceBlack                   0.119855   0.209716   0.572
0.56766
## RaceOther                  -0.011988   0.387597
-0.031  0.97533
## RaceWhite                   0.537955   0.200711
2.680  0.00736 **
## SexMale                     1.191472   0.041274  28.868
< 2e-16 ***
## `Hours-per-Week`            0.033270   0.001458
22.814  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
##
## (Dispersion parameter for binomial family taken
to be 1)
##
##     Null deviance: 31191  on 27502  degrees of
freedom
## Residual deviance: 24111  on 27474  degrees of
freedom
## AIC: 24169
##
## Number of Fisher Scoring iterations: 12
drop1(log.model,test = "Chisq")
## Single term deletions
##
## Model:
## Income ~ Age + Workclass + Education +
`Marital-Status` + Occupation +
##     Relationship + Race + Sex + `Hours-per-Week`
##                 Df Deviance   AIC     LRT Pr(>Chi)
## <none>               19845 19951
## Age              1    20143 20247  297.89   <2e-16
***
## Workclass        6    19981 20075  135.56   <2e-16
***
## Education       15    21004 21080 1158.79   <2e-16
***
## `Marital-Status`  6    19940 20034   95.00   <2e-16
***
## Occupation      13    20460 20540  614.33
<2e-16 ***
## Relationship     5    20147 20243  302.14   <2e-16
***
## Race             4    19857 19955   11.81   0.0188 *
## Sex              1    19977 20081  131.80   <2e-16
***
## `Hours-per-Week`  1    20164 20268  318.25
<2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
#Coefficient of the final model
coef(final.model)
##            (Intercept)                    Age
WorkclassLocal-gov
##           -6.96356153             0.04729158
-0.47216845
##       WorkclassPrivate  WorkclassSelf-emp-inc
WorkclassSelf-emp-not-inc
##           -0.39560328             0.10621827
-0.87665564
##      WorkclassState-gov
WorkclassWithout-pay           Education11th
##           -0.62323858           -12.83720345
0.06879276
##         Education12th         Education1st-4th
Education5th-6th
##            0.47128651            -1.49080872
-1.27778840
##       Education7th-8th           Education9th
EducationAssoc-acdm
##           -0.64025889            -0.33230786
1.73543646
##      EducationAssoc-voc      EducationBachelors
EducationDoctorate
##            1.68023367             2.40342888
3.48477555
##       EducationHS-grad       EducationMasters
EducationPreschool
##            0.99968050             2.85101675
-10.64007772
##    EducationProf-school
EducationSome-college  RaceAsian-Pac-Islander
##            3.46645295             1.40948658
0.45855625
##             RaceBlack              RaceOther
RaceWhite
##            0.11985456            -0.01198791
0.53795468
##               SexMale        `Hours-per-Week`
##            1.19147195             0.03326985
#Equation for the final model
paste("logit(p) = ", round(coef(final.model)[1], 4),
"+",
    paste(round(coef(final.model)[-1], 4),
names(coef(final.model)[-1]), collapse = " + "))
#Plotting the final model
```

```
adult.data$prob <- predict(final.model, type =
"response")
#Plot of predicted values vs. actual values
ggplot(adult.data, aes(x = Income, y = prob)) +
  geom_smooth(method = "glm", method.args =
list(family = "binomial"), se = FALSE) +
  geom_point(alpha = 0.3) +
  labs(title = "Logistic Regression Plot", x = "Actual
Income", y = "Predicted Probability")
## `geom_smooth()` using formula 'y ~ x'
## Warning in eval(family$initialize): non-integer
#successes in a binomial glm!
  #Checking accuracy of model against other models
#Final model
pred_probs <- predict(final.model, type = "response")
pred_class <- ifelse(pred_probs > 0.5, "Yes", "No")
# Create a confusion matrix
conf_mat <- table(Predicted = pred_class, Actual =
adult.data$Income)
conf_mat
##          Actual
## Predicted     0     1
##       No  19116  4192
##       Yes  1392  2803
accuracy.final <- sum(diag(conf_mat)) /
sum(conf_mat)
accuracy.final
## [1] 0.7969676
#compared to all of them including martial and
relationship
#Compared to Using them all(except for marital
status and relationship because of multicollinearity.)
basically i put occupation back
#the main point here is that they predict it better but i
don't know if all those variables would cause
overfitting in the future. I feel that the model is
complicated enough w/ education but I didn't want to
disinclude it becasue it was the main factor of my
hypothesis.
log.model1 <-
glm(Income~Age+Workclass+Education+Occupation
+Race+Sex+`Hours-per-Week`, data = adult.data,
family = "binomial")
pred_probs1 <- predict(log.model1, type =
"response")
pred_class1 <- ifelse(pred_probs1 > 0.5, "Yes", "No")
# Create a confusion matrix
conf_mat1 <- table(Predicted = pred_class1, Actual =
adult.data$Income)
conf_mat1
##          Actual
## Predicted     0     1
##       No  19020  3878
##       Yes  1488  3117
accuracy.final1 <- sum(diag(conf_mat1)) /
sum(conf_mat1)
accuracy.final1
## [1] 0.804894
#Finally I wanted to compare the final model I
created with a model that would not Include the
Education variable(had the largest AIC).
log.model2 <-
glm(Income~Age+Workclass+Race+Sex+`Hours-per
-Week`, data = adult.data, family = "binomial")
pred_probs2 <- predict(log.model2, type =
"response")
pred_class2 <- ifelse(pred_probs2 > 0.5, "Yes", "No")
# Create a confusion matrix
conf_mat2 <- table(Predicted = pred_class2, Actual =
adult.data$Income)
conf_mat2
##          Actual
## Predicted     0     1
##       No  19308  5606
##       Yes  1200  1389
accuracy.final2 <- sum(diag(conf_mat2)) /
sum(conf_mat2)
accuracy.final2
## [1] 0.7525361
```