Data Analytics II Final Project

Ronnald Le

2023-06-13

Datsets:

1. Census income: https://archive.ics.uci.edu/ml/datasets/Adult Potential responses: whether or not making more than 50k; whether or not completed college; etc.

Data Description:

The Census Income data set provides valuable information about individuals and its attributes is allowing us to gain insights into their income levels. The data set includes information such as: age, education, work class, marital status, occupation, and more. Two potential responses of interest that caught my eye is in this data set are whether an individual earns more than $50k per year and whether they have completed college (potential response mentioned above).

This data set has a substantial number of observations, with each observation representing a unique individual. It contains a combination of categorical and numerical variables, offering a comprehensive view of the individuals characteristics.

The dataset "census" was taken from the https://archive.ics.uci.edu/dataset/2/adult website, where it was downloaded into a zip file called "adult.zip". The zip file was uploaded to R Studio Cloud (Posit Cloud). The column names were labeled as V1 to V15, which was then corrected in Question 1.

Variable & Analysis Methods:

For the first question, we can use logistic regression to model the probability of an individual's income being greater than $50k based on the explanatory variables. The investigation aimed to identify the factors associated with an individual earning more than $50,000 per year. Logistic regression was employed to model the probability of income exceeding this threshold, considering variables such as age, education level, occupation, work class, and marital status.

For the second question, we can conduct a hypothesis test using a chi-square test or logistic regression to determine the association between completing college and income level. The study examined whether completing college has a significant impact on an individual's income level. A chi-square test of independence was conducted, analyzing the association between completing college ("Yes" or "No") and income level (">50K" or "<=50K").

Questions:

1.What factors are associated with an individual making more than $50,000 per year?

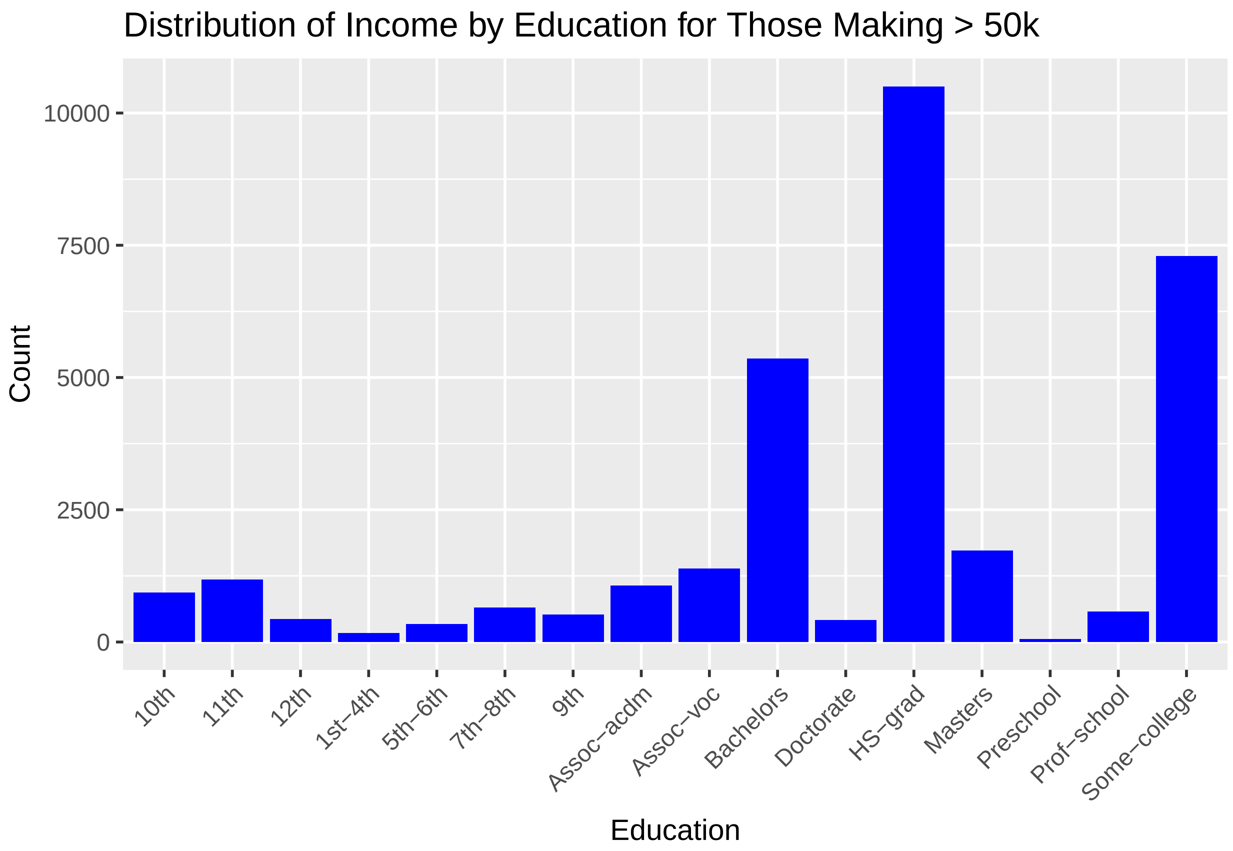
Response variable: Income (>50K or <=50K)

```r
# Assign appropriate column names to the data set due to dataset not properly labeling
colnames(census) <- c("age", "state.gov", "blank", "education", "workclass",
                      "marital_status", "occupation",
                      "relationship", "race", "sex",
                      "capital_gain", "capital_loss",
                      "hours_per_week", "native_country",
                      "income")

# Verify the updated column names
colnames(census)
```

Explanatory variables: education, race, hours per week

```
##  [1] "age"            "state.gov"      "blank"          "education"
##  [5] "workclass"      "marital_status" "occupation"     "relationship"
##  [9] "race"           "sex"            "capital_gain"   "capital_loss"
## [13] "hours_per_week" "native_country" "income"
```

```r
# Create a bar plot of education for those making > $50,000 per year
ggplot(census, aes(x = education, fill = as.character(income > 50000))) +
  geom_bar(fill = c("TRUE" = "blue")) +
  labs(x = "Education", y = "Count", fill = "Income > 50K") +
  ggtitle("Distribution of Income by Education for Those Making > 50k") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The logistic regression model was fitted to investigate the relationship between income and the variables education, race, and hours per week. The p-values for all variables are extremely high (close to 1 or is 1), which indicates that none of the variables are statistically significant in predicting income. The null deviance is zero, and the residual deviance is also close to zero, suggesting a good fit of the model to the data. Based on the given variables (education, race, and hours per week), there is no statistically significant evidence to conclude that any particular factor is associated with income levels above $50,000. Based on the bar plot, the likeness of those who are HS-grad without any further higher education makes more than those who do have a higher education.

2. Does completing college have a significant impact on an individual's income level?

Response variable: Income (>50K or <=50K)

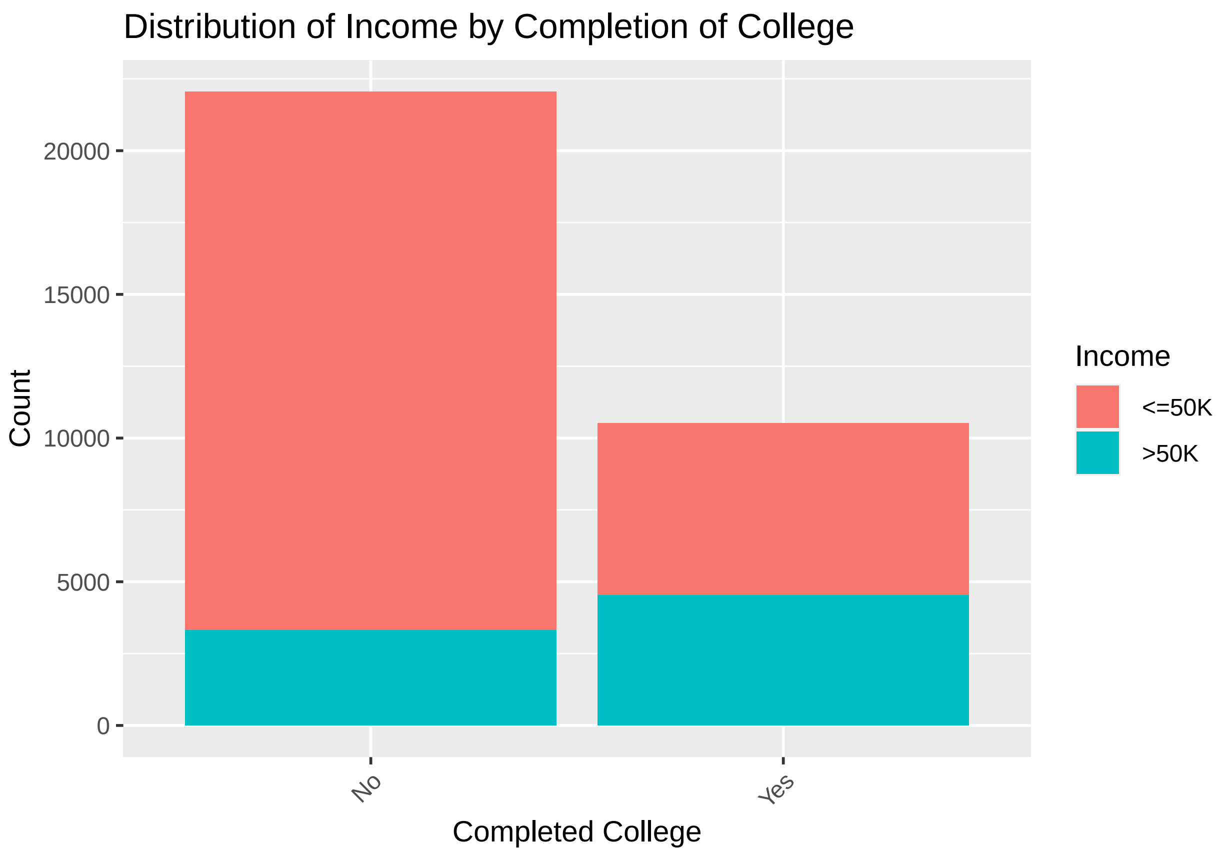
```r
# Define the levels for completed college
completed_college_levels <- c(" Assoc-acdm", " Assoc-voc", " Bachelors",
                              " Doctorate", " Masters", " Prof-school")

# Create a new variable indicating completion of college
census$Completed_college <- ifelse(census$education %in% completed_college_levels,
                                   "Yes", "No")
```

```r
# Create a contingency table of Completed_college and Income
contingency_table <- table(census$Completed_college, census$income)

# Perform a chi-square test of independence
chi_sq_test <- chisq.test(contingency_table)

# Results of the chi-square test
chi_sq_test
```

Explanatory variable: Completed college (Yes or No)

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 3079.7, df = 1, p-value < 2.2e-16
```

```r
# Create the bar plot
ggplot(census, aes(x = Completed_college, fill = income)) +
  geom_bar() +
  labs(x = "Completed College", y = "Count", fill = "Income") +
  ggtitle("Distribution of Income by Completion of College") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The chi-square test of independence was conducted to examine the association between completing college and an individual's income level. The analysis revealed a significant relationship between these two variables chi-sq = 3079.7, df = 1, and p < 2.2e-16. The large chi-square test statistic suggests a substantial discrepancy between the observed frequencies in the contingency table and the expected frequencies under the assumption of independence. However, the small p-value shows strong evidence against the null hypothesis, which indicates that completing college has a significant impact on an individual's income level. Based on the question of interest, education plays a crucial role in determining income outcomes, with completing college being associated with a higher likelihood of achieving a higher income level.

Conclusion

The logistic regression model examined the relationship between income and variables such as education, race, and hours per week. The results revealed no statistically significant evidence to conclude that any specific factor is associated with income levels above $50,000. This finding challenges the commonly held notion that education, race, or hours per week alone can reliably predict higher income outcomes. On the other hand, the chi-square test of independence demonstrated a significant association between completing college and income level. This finding suggests that completing college has a substantial impact on an individual's income, with a higher likelihood of achieving a higher income level for those who have completed college. The significant association between completing college and income level reinforces the value of higher education as a pathway to increased earning potential shown the in second question.

Appendices

```r
# 2. Fit logistic regression model
model <- glm(income > 50000 ~ education + race + hours_per_week, data = census,
             family = "binomial")

# 3. Summary of the logistic regression model
summary(model)
```

```
##
## Call:
## glm(formula = income > 50000 ~ education + race + hours_per_week,
##     family = "binomial", data = census)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06  -2.409e-06
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -2.657e+01  2.390e+04  -0.001    0.999
## education 11th          -2.063e-14  1.563e+04   0.000    1.000
## education 12th           3.508e-15  2.072e+04   0.000    1.000
## education 1st-4th        9.637e-15  2.987e+04   0.000    1.000
## education 5th-6th        8.384e-15  2.275e+04   0.000    1.000
## education 7th-8th        3.735e-15  1.824e+04   0.000    1.000
## education 9th            9.458e-15  1.956e+04   0.000    1.000
## education Assoc-acdm     9.291e-15  1.598e+04   0.000    1.000
## education Assoc-voc      9.307e-15  1.511e+04   0.000    1.000
## education Bachelors      1.426e-12  1.269e+04   0.000    1.000
## education Doctorate      3.796e-14  2.113e+04   0.000    1.000
## education HS-grad        9.553e-15  1.218e+04   0.000    1.000
## education Masters        3.221e-14  1.454e+04   0.000    1.000
## education Preschool      3.908e-14  5.123e+04   0.000    1.000
## education Prof-school    4.205e-14  1.897e+04   0.000    1.000
## education Some-college  -6.581e-15  1.239e+04   0.000    1.000
## race Asian-Pac-Islander -3.013e-13  2.308e+04   0.000    1.000
## race Black              -6.390e-14  2.118e+04   0.000    1.000
## race Other              -8.815e-14  2.963e+04   0.000    1.000
## race White               1.193e-13  2.032e+04   0.000    1.000
## hours_per_week          -4.406e-15  1.631e+02   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 0.0000e+00  on 32560  degrees of freedom
## Residual deviance: 1.8891e-07  on 32540  degrees of freedom
## AIC: 42
##
## Number of Fisher Scoring iterations: 25
```