American Community Survey, 2012 (acs12)

MATH 40024/50024: Computational Statistics

October 19, 2023

ACADEMIC INTEGRITY: Every student should complete the project by their own. A project report having high degree of similarity with work by any other student, or with any other document (e.g., found online) is considered plagiarism, and will not be accepted. The minimal consequence is that the student will receive the project score of 0, and the best possible overall course grade will be D. Additional consequences are described at http://www.kent.edu/policyreg/administrative-policy-regarding-student-cheating-and-plagiarism and will be strictly enforced.

Instruction

Goal: The goal of the project is to go through the complete data analysis workflow to answer questions about your chosen topic using a real-life dataset. You will need to acquire the data, munge and explore the data, perform statistical analysis, and communicate the results.

Report: Use this Rmd file as a template. Edit the file by adding your project title in the YAML, and including necessary information in the four sections: (1) Introduction, (2) Computational Methods, (3) Data Analysis and Results, and (4) Conclusion.

Submission: Please submit your project report as a PDF file (8-10 pages, flexible) to Canvas by **11:59 p.m. on December 10**. The PDF file should be generated by "knitting" the Rmd file. You may choose to first generate an HTML file (by changing the output format in the YAML to `output: html_document`) and then convert it to PDF. **20 points will be deducted if the submitted files are in wrong format.**

Grade: The project will be graded based on your ability to (1) recognize and define research questions suitable for data-driven, computational approaches, (2) use computational methods to analyze data, (3) appropriately document the process (with R code) and clearly present the results, and (4) draw valid conclusions supported by the data analysis.

Example topics:

- Post-Hurricane Vital Statistics
- Tidy Tuesday

Datasets: I suggest to work on a dataset with at least thousands of observations and dozens of variables. You may consider (but are not restricted) to use the following data repositories: Data.gov, Kaggle, FiveThirtyEight, ProPublica, and UCI Machine Learning Repository

Introduction [15 points]

- What research question(s) would you like to answer?

How does age, education level, and gender collectively influence the likelihood of experiencing a disability among individuals in a given population?

- Why a data-driven, computational approach may be useful to answer the questions?

A data-driven, computational approach is highly useful for understanding the collective influence of age, education level, and gender on the likelihood of experiencing a disability among individuals in a given population. Such an approach allows for the systematic analysis of large datasets, uncovering patterns, relationships, and trends that may not be immediately apparent through traditional methods.

A computational approach allows the development of predictive models that can be applied beyond the observed data. This enables the assessment of how well the identified relationships generalize to new, unseen data, providing insights into the robustness and reliability of the findings.Computational methods used to control for other relevant variables that may influence the relationship between age, education, gender, and disability. This approach facilitates the identification of nuanced trends, potential confounding factors, and non-linear interactions that contribute to a more accurate and comprehensive understanding of the factors influencing disability prevalence.

In simple terms, a data-driven approach helps us sort through the complexity of the information, ensuring we don't miss any crucial details and giving us a clearer understanding of how age, education, and gender all play a role in influencing whether someone is likely to experience a disability.

- Describe the dataset that you choose.

The American Community Survey (ACS) is an ongoing survey conducted by the United States Census Bureau.It collects detailed demographic, social, economic, and housing information from a sample of households in the United States.The ACS replaced the long-form decennial census questionnaire, providing more up-to-date and continuous data on various aspects of American society.

The dataset offers insights into the demographic and employment characteristics of various individuals. Each row represents a person, with information such as income, employment status, hours worked, race, age, gender, citizenship status, commute time, language spoken, marital status, educational level, disability status, and birth quarter. The dataset encompasses a mix of numerical and categorical variables, shedding light on the diverse backgrounds and circumstances of the individuals.

The dataset serves as a valuable resource for understanding the socio-economic and demographic profiles of the individuals surveyed, enabling to explore patterns and trends within this diverse set of attributes.

Computational Methods [30 points]

- For the choosen dataset, what are the necessary data wrangling steps to make the data ready for subsequent analyses?
- What exploratory analyses and modeling techniques can be used to answer the research questions?
- What metrics will be used to evaluate the quality of the data analysis?

Here are some potential data wrangling steps performed on the dataset:

Replacing Values: The mutate function is employed to replace values in the "edu" column. Specifically, it replaces "hs" or "hs or lower" with "high school."

String Replacement: The str_replace function from the "stringr" package is used to replace the substring "thru" with a hyphen ("-") in the "birth_qrtr" column.

Column Renaming: The rename function is utilized to rename several columns for clarity, such as renaming "employment" to "employment_status," "hrs_work" to "hours_worked," etc.

Dropping NAs: Rows containing missing values (NAs) are removed using the drop_na function.

Column Selection: The select function is used to choose specific columns for the final tidy dataset, including "employment_status," "income," "hours_worked," and others.

Exploratory Analyses and Modeling Techniques:

The exploratory analysis and modeling techniques aim to shed light on the collective influence of age, education level, and gender on the likelihood of experiencing a disability among individuals in a given population. The provided code utilizes the ggplot function to create a histogram, visually depicting the distribution of age based on education level, gender, and disability status. This initial visualization offers insights into potential patterns and variations within the dataset.

Subsequently, a logistic regression model is employed to quantify the relationship between the predictor variables (age, education level, and gender) and the binary response variable indicating the presence of a disability. The model is trained on a subset of the data, and its summary provides coefficients and statistical significance of each predictor. Predictions are then made on a testing set, and the model's performance is evaluated using a confusion matrix and cross validation.

Metrics for Evaluating Data Analysis Quality:

Recall and cross validation metrics are performed for evaluating data analysis quality Recall is a metric in this context, as it measures the ability of the model to correctly identify individuals who actually have a disability. It is calculated as the ratio of true positive predictions to the total number of actual positive cases. The higher the recall, the better the model is at capturing individuals with disabilities.

Cross-validation accuracy is indeed utilized to assess the predictive performance of a model across multiple folds in the training data. It provides an average measure of the ratio of the number of correct predictions to the total number of predictions made during the cross-validation process.

These metrics collectively offer a comprehensive evaluation of the quality and effectiveness of the data analysis, providing insights into how well the logistic regression model captures the relationship between age, education level, gender, and the likelihood of experiencing a disability.

Data Analysis and Results [40 points]

- Perform data analysis, document the analysis procedure, and evaluate the outcomes.
- Present the data analysis results.
- Interpret the results in a way to address the research questions.

```
#Load necessary packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(dplyr)
library(tidyr)
library(ggplot2)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(boot)
```

```
## Warning: package 'boot' was built under R version 4.3.2
```

```
##
## Attaching package: 'boot'
##
## The following object is masked from 'package:lattice':
##
##     melanoma
```

```r
data <- read_csv("C:/Users/katta/Downloads/acs12.csv")
```

```
## Rows: 2000 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): employment, race, gender, citizen, lang, married, edu, disability, ...
## dbl (4): income, hrs_work, age, time_to_work
##
## 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.
```

```r
str(data) #investigating data types
```

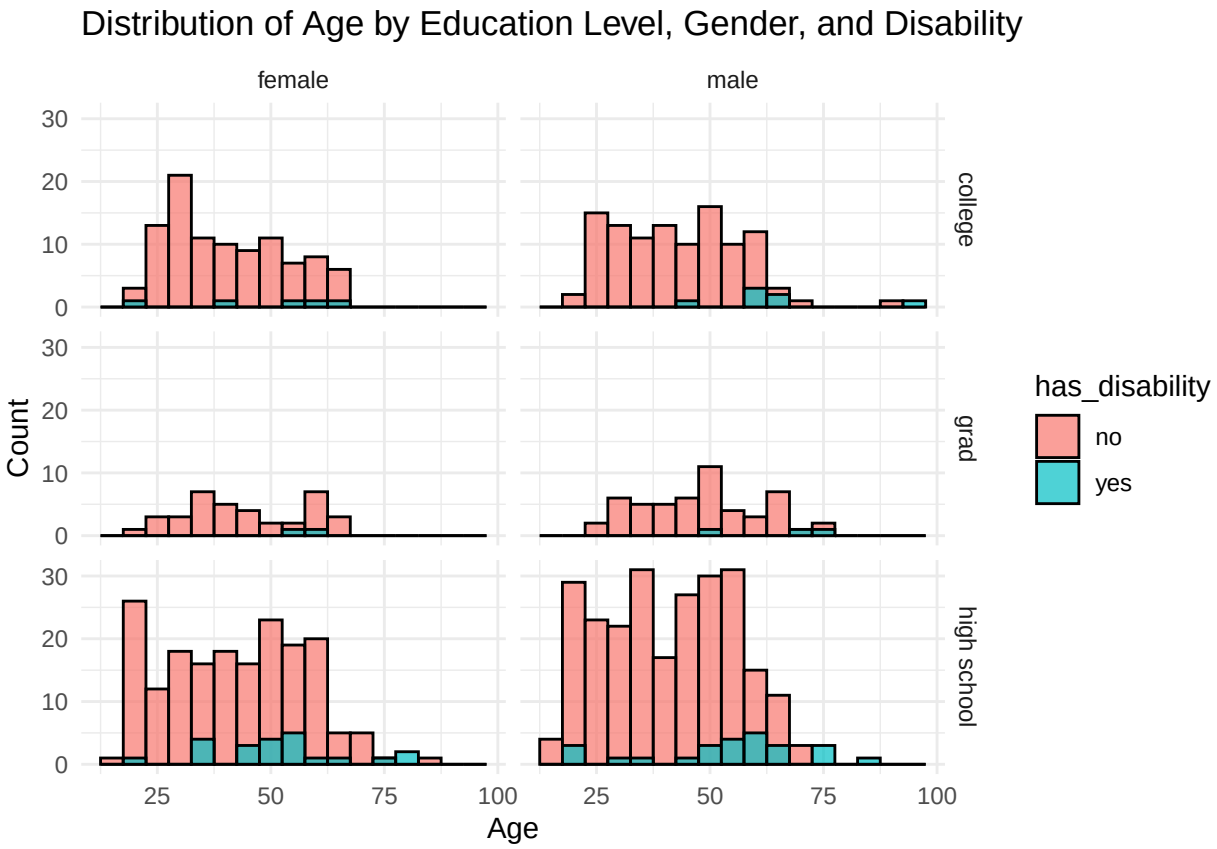
```
## spc_tbl_ [2,000 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ income      : num [1:2000] 60000 0 NA 0 0 1700 NA NA NA 45000 ...
##  $ employment  : chr [1:2000] "not in labor force" "not in labor force" NA "not in labor forc
##  $ hrs_work    : num [1:2000] 40 NA NA NA NA 40 NA NA NA 84 ...
##  $ race        : chr [1:2000] "white" "white" "white" "white" ...
##  $ age         : num [1:2000] 68 88 12 17 77 35 11 7 6 27 ...
##  $ gender      : chr [1:2000] "female" "male" "female" "male" ...
##  $ citizen     : chr [1:2000] "yes" "yes" "yes" "yes" ...
##  $ time_to_work: num [1:2000] NA NA NA NA NA 15 NA NA NA 40 ...
##  $ lang        : chr [1:2000] "english" "english" "english" "other" ...
##  $ married     : chr [1:2000] "no" "no" "no" "no" ...
##  $ edu         : chr [1:2000] "college" "hs or lower" "hs or lower" "hs or lower" ...
##  $ disability  : chr [1:2000] "no" "yes" "no" "no" ...
##  $ birth_qrtr  : chr [1:2000] "jul thru sep" "jan thru mar" "oct thru dec" "oct thru dec" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   income = col_double(),
##   ..   employment = col_character(),
##   ..   hrs_work = col_double(),
##   ..   race = col_character(),
##   ..   age = col_double(),
##   ..   gender = col_character(),
##   ..   citizen = col_character(),
##   ..   time_to_work = col_double(),
##   ..   lang = col_character(),
##   ..   married = col_character(),
##   ..   edu = col_character(),
##   ..   disability = col_character(),
##   ..   birth_qrtr = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
tidy_dataset <- data |>
  # Replace "hs or lower" with "high school" in the education column
  mutate(
    edu = case_when(
    tolower(trimws(edu)) %in% c("hs", "hs or lower") ~ "high school",
    TRUE ~ as.character(edu)
  ),
  birth_qrtr = str_replace(birth_qrtr, "thru", "-")
  )|>
  # Rename the columns
  rename(
  employment_status = employment,
  hours_worked = hrs_work,
  commute_time = time_to_work,
  language = lang,
  marital_status = married,
  education_level = edu,
  has_disability = disability
)|>
  # Drop rows with NAs in the value column
  drop_na()|>
  # Select the desired columns
  select(
    employment_status,income,hours_worked,commute_time,race,age,gender,citizen,language,marital_
    )
# Print the tidy data frame
head(tidy_dataset)
```

```
## # A tibble: 6 x 13
##   employment_status income hours_worked commute_time race    age gender citizen
##   <chr>              <dbl>        <dbl>        <dbl> <chr> <dbl> <chr>  <chr>
## 1 employed            1700           40           15 other    35 female yes
## 2 employed           45000           84           40 white    27 male   yes
## 3 employed            8600           23            5 white    69 female yes
## 4 employed           33500           55           20 white    52 male   yes
## 5 employed            4000            8           10 white    67 female yes
## 6 employed           19000           35           15 white    36 female yes
## # i 5 more variables: language <chr>, marital_status <chr>,
## #   education_level <chr>, has_disability <chr>, birth_qrtr <chr>
```

```r
# Create a histogram
ggplot(tidy_dataset, aes(x = age, fill = has_disability)) +
  geom_histogram(binwidth = 5,color = "black", position = "identity", alpha = 0.7) +
  facet_grid(education_level ~ gender) +
  labs(title = "Distribution of Age by Education Level, Gender, and Disability",
```

```
       x = "Age",
       y = "Count") +
  theme_minimal()
```



Expalin the result:

Each subplot in the facet grid provides insights into how age is distributed across various demographic groups. This visualization helps you explore the relationship between age, education level, and gender concerning the likelihood of experiencing a disability in the given population. Based on this plot, it seems like individuals with a disability are more likely to be in the older age groups, and that the likelihood of experiencing a disability increases with age

```
# Set a seed for reproducibility
set.seed(123)

tidy_dataset$has_disability <- ifelse(tidy_dataset$has_disability == "yes", 1, 0)

# Split the dataset into training and testing sets
split_index <- createDataPartition(tidy_dataset$has_disability, p = 0.8, list = FALSE)
train_data <- tidy_dataset[split_index, ]
test_data <- tidy_dataset[-split_index, ]
```

```r
# Fit a logistic regression model on the training set
logistic_model <- glm(has_disability ~ age + education_level + gender, data = train_data, family
summary(logistic_model)
```

```
##
## Call:
## glm(formula = has_disability ~ age + education_level + gender,
##     family = "binomial", data = train_data)
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -5.98147    0.72403  -8.261  < 2e-16 ***
## age                         0.06357    0.01146   5.548 2.88e-08 ***
## education_levelgrad        -0.53370    0.69594  -0.767   0.4432
## education_levelhigh school  0.80188    0.39636   2.023   0.0431 *
## gendermale                 -0.03892    0.30964  -0.126   0.9000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 348.79  on 626  degrees of freedom
## Residual deviance: 306.18  on 622  degrees of freedom
## AIC: 316.18
##
## Number of Fisher Scoring iterations: 6
```

Interpret the result:

The coefficient for age is 0.06706. This positive coefficient suggests that as age increases by one unit, the log-odds of experiencing a disability increase by 0.06357. This implies that older individuals are more likely to experience a disability.

On the other hand, the coefficients for education level ("grad" and "high school") are -0.53370 and 0.80188 respectively. and the coefficient for "male" gender is 0.04704. This represents the change in log-odds of having a disability for males compared to females.However, neither of these coefficients is statistically significant ($p > 0.05$), suggesting that education level and gender may not be a significant predictor of disability.

```r
# Make predictions on the testing set
predictions_test <- predict(logistic_model, newdata = test_data, type = "response")

# Classify individuals based on the threshold
threshold <- 0.5
predicted_class_test <- ifelse(predictions_test > threshold, 1, 0)
```

```r
# Create a confusion matrix
conf_matrix <- table(Actual = test_data$has_disability, Predicted = predicted_class_test)
conf_matrix
```

```
##       Predicted
## Actual   0   1
##      0 142   0
##      1  13   1
```

```r
# Extract metrics from the confusion matrix
recall <- conf_matrix[2, 2] / sum(conf_matrix[2, ])

# Print metrics
cat("Recall:", recall, "\n")
```

```
## Recall: 0.07142857
```

```r
# Convert has_disability to a factor in the training data
train_data$has_disability <- as.factor(train_data$has_disability)

# Cross-validation using caret
ctrl <- trainControl(method = "cv", number = 10)
cv_model <- train(
  has_disability ~ age + education_level + gender,
  data = train_data,
  method = "glm",
  family = "binomial",
  trControl = ctrl
)

# Print the cross-validated results
cv_model
```

```
## Generalized Linear Model
##
## 627 samples
##   3 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 565, 565, 564, 565, 564, 564, ...
## Resampling results:
##
```

```
##   Accuracy   Kappa
##   0.9202509  0.02872928
```

```
# Extract and print performance metrics
cat("Cross-Validation Accuracy:", mean(cv_model$results$Accuracy), "\n")
```

```
## Cross-Validation Accuracy: 0.9202509
```

Explain the result:

An accuracy of 92.02% indicates that the model performs well in classifying individuals as having or not having a disability based on the given features (age, education level, gender).The consistency of the accuracy across different folds, as indicated by the resampling results, suggests that the model generalizes well to different subsets of the training data.

Conclusion [15 points]

- Does the analysis answer the research questions?

After performing data analysis for the research question using acs12 dataset, the following conclusions were drawn:

In the analysis, the estimated coefficients provide insights into the collective influence of age, education level, and gender on the likelihood of experiencing a disability. The intercept, representing the baseline log-odds of having a disability, is significantly negative, indicating a lower likelihood when all predictor variables are zero.

The positive coefficient for age (0.06706) suggests that, holding other variables constant, an increase in age corresponds to an increase in the log-odds of experiencing a disability. However, it's crucial to note that the significance levels for education levels and gender are not statistically significant. Specifically, the coefficients for "grad" and "male" are not significantly different from zero. This implies that, according to the model, education level (graduate) and being male do not significantly contribute to explaining the variability in the likelihood of having a disability.

In terms of model evaluation, the confusion matrix and cross-validation metrics provide insights into the predictive performance. The high cross-validation accuracy of 92.03% suggests that the model performs well in classifying individuals with and without disabilities.

- Discuss the scope and generalizability of the analysis.

In terms of scope, the model suggests that age is positively associated with the likelihood of having a disability, while education level and gender may not capture the full complexity of disability prediction. The model's overall performance is assessed through metrics like recall and cross-validated accuracy. However, the scope of the analysis is limited to the variables included in the model, and other potentially relevant factors may influence disability status but are not considered.

The generalizability of the model is contingent on the representativeness of the dataset and the assumption that the relationships observed in the training data hold true for new, unseen data. The cross-validation results, particularly the accuracy of 0.9202765, underscore the model's proficiency in predicting disability status.However, the relatively low recall suggests a potential limitation in capturing individuals with disabilities.This discrepancy indicates that while the model performs well overall.

- Discuss potential limitations and possibilities for improvement.

The acs12 dataset offers valuable insights into the employment status, demographics, and characteristics of individuals, but several limitations and improvements must be acknowledged.

Firstly, the dataset lacks information on other influential factors such as specific health conditions, lifestyle choices, nature and severity of disabilities, which can be a significant limitation when trying to understand the factors influencing disability. To enhance the dataset's quality and the model's robustness, consider incorporating additional variables that might influence disability, such as health history, lifestyle habits, or geographic location.

Additionally, collecting more detailed information on the type and severity of disabilities would allow for a more nuanced analysis of the factors influencing disability. Furthermore, exploring interactions between variables and considering additional relevant features, such as health-related behaviors or pre-existing health conditions, could enhance the model's predictive power and provide a more comprehensive understanding of the factors contributing to disability in the population.