Income Influences: A Logistic Regression Analysis

YESSICA RUBIO

MAY 29, 2025

Contents

- Overview
- Executive summary
- Introduction
- Analysis questions
- Methods
- Results
- Scope of inference/Limitations
- Conclusion/Discussion

Overview

We will look at valuable insights with logistic regression to uncover how income is influenced by factors such as education, age, gender, and hours worked. Understanding these relationships is essential for shaping policies and practices related to employment equity, workforce development, and addressing income inequality.

Executive Summary

This analysis investigates the relationship between demographic and employment factors using 1994 U.S. Census data.

- Understand the factors that influence whether individuals earn more than $50,000 annually.
- Logistic regression models were applied to a dataset from the UCI Machine Learning Repository.

(1) Is there an association between earning over $50,000 and hours worked per week, after accounting for other factors?

- After adjusting for other factors, for each additional hour worked per week, the odds of earning over $50,000 was 1.03 times the odds for someone who works one hour less.

(2) After accounting for age, is there evidence of a difference in the earnings probability of men and women?

- The odds ratio for men compared to women was 1.02, meaning that men had 1.02 times the odds of earning over $50,000 than women of the same age.

- Analysis supports decisions that promote fairness and efficiency in the labor market.
- Work hours and gender disparities should be considered in income-related workforce planning.
- Data from 1994 is outdated and may not reflect today’s dynamics but results still support data-driven efforts to improve fairness in employment practices.

Introduction

Data Description

- The 1994 U.S. Census dataset hosted by the UCI Machine Learning Repository was used in this analysis.
- Primary objective: If an individual had an income that exceeded $50,000 per year based on the association with specific attributes.
- Processed by Barry Becker: consists of 32,561 observations and 15 variables (earnings50k is binary).
- Missing or unknown values were removed from workclass, occupation, and native-country
    - Ensure a clean data set for modeling and maintain data integrity
    - Prevent noise or bias through imputation.
- Education-num: a numeric encoding of the variable 'education'. This would introduce redundancy and multicollinearity.

Data Variables:

Age
Workclass
fnlwgt
Education
Education-num
Marital-status
Occupation
Relationship
Race
Sex
Capital-gain
Capital-loss
Hours-per-week
Native-country
Earnings50k - (Binary response variable, indicating whether an individual earns more than $50,000 annually or not.

Methods

Exploratory

Explored the relationship between earnings and explanatory variables using summary statistics and visualizations to guide model building. Checked key logistic regression assumptions.

Hours Worked

Question of interest 1: Fit logistic regression models, first without and the second with the hours per week variable, to assess its contribution.

Interaction Term

Question of interest 2: Fit logistic regression models, one with age, sex, and earnings50k to isolate and examine differences in earnings probability between men and women, and a second that included an interaction term between sex and age.

Model fit

Use Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) from likelihood ratio test to compare models, as it offers a robust way to assess fit while balancing model complexity. In logistic regression, residuals aren't normally distributed and are less informative.

Why Logistic Regression?

STATISTICAL MODEL USED TO PREDICT THE PROBABILITY OF A BINARY OUTCOME.

THE RESPONSE VARIABLE **'EARNINGS50K'** IN THIS DATASET IS BINARY (>$50K OR ≤$50K).

IT ALLOWS FOR THE ESTIMATION OF ODDS RATIOS.

How does it work?

Predicts the probability of an outcome through something called log-odds.

- This unique transformation function, Logit function to keep predictions between the binary outcomes.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

p is the probability of earning >50k and p/1-p is the odds.

Interpretation

- Each **coefficient** βj tells us how the **log-odds** of the outcome change when the predictor increases by one unit.
- Since log-odds aren't intuitive, we usually **exponentiate** the coefficient to get something called an **odds ratio**.

Assumptions

Binary Outcome
The target variable (earnings50k) is binary (<=50K vs. >50K).

Independent Observations
Each record is an individual assumed to be independent of the others.

No Perfect Multicollinearity
Predictors (e.g., education, education-num) are linearly dependent. Only education will be used in the analysis.

Sufficient Sample Size
With over 32,000 records, the sample size is more than adequate.

Analysis Questions

There are two main questions of interest guiding the analysis:

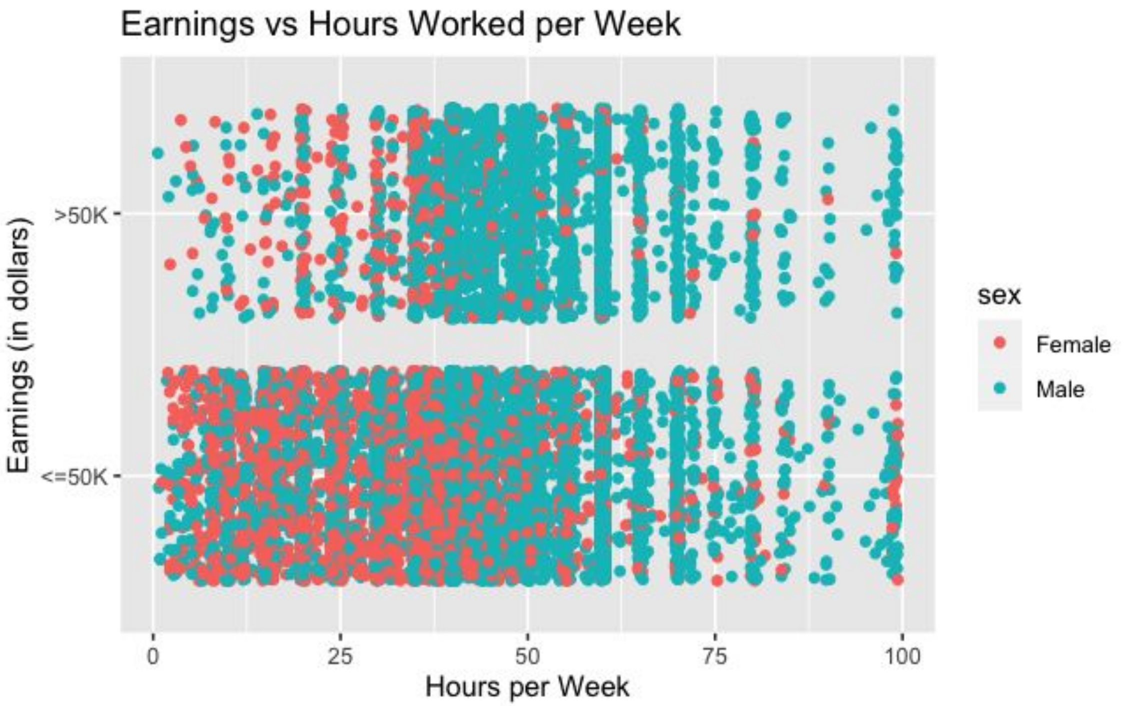
- First, is there an association between earnings over $50,000 and hours worked per week, after accounting for other factors (sex, age, education, workclass, final-weight, marital-status, occupation, relationship, race, capital-gain, capital-loss, native-country)?

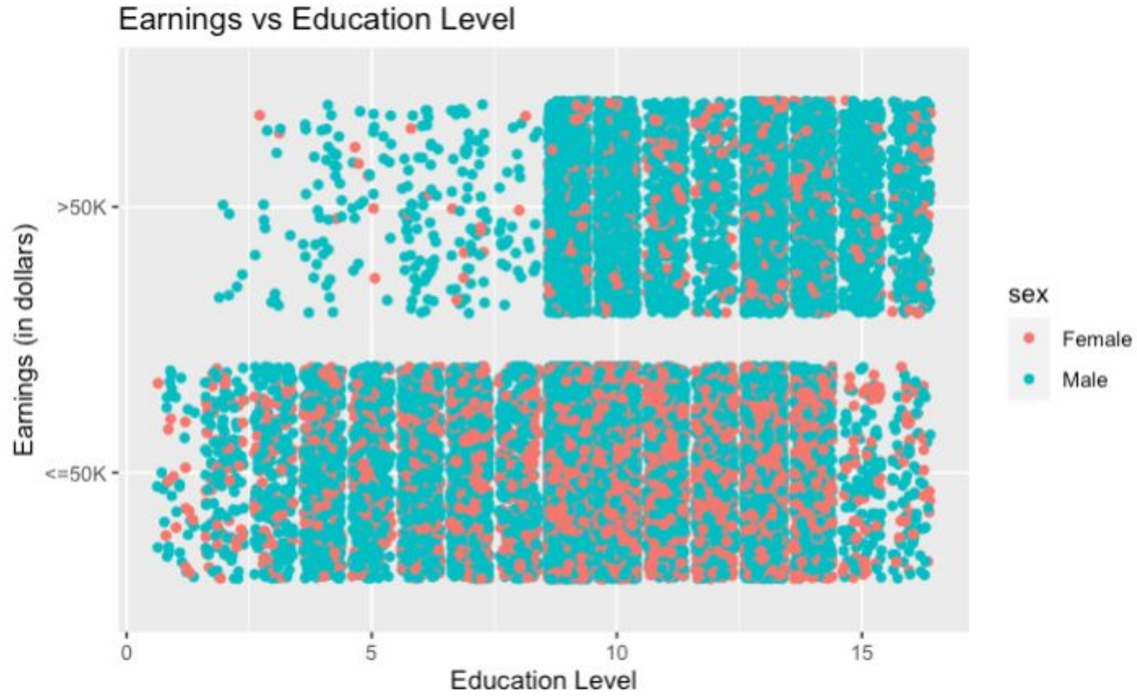
- Second, after accounting for age, is there evidence of a difference in the earnings probability of men and women?

Results

Exploratory Data Analysis

- Plots have binary outcomes (>50k and <=50k), suitable for logistic regression.
- Most of the individuals in the dataset earned $50,000 or less as seen in the scatterplots
- Males, on average, tend to work more hours per week than females.
- Visible concentration around 40 hours worked per week among both genders in the first plot and a possible trend where individuals with more years of education could be more likely to make more than $50,000 in the second plot.
- There are more individuals with a higher education level that earn more than $50,000.





Results

Earnings and Hours worked

- Model 1: Fit logistic regression model without 'hours-per-week' variable.
  - Had significant predictor variables, such as education, occupation, and workclass.

- Model 2: Fit logistic regression model including 'hours-per-week' variable.
  - The addition of the variable ‘hours-per-week' added a strong effect (p-value < 0.001) to the model with a coefficient of 0.02949,translating to an odds ratio of 1.03.
  - Interpretation: There is a significant association between earning over $50,000 annually and the number of hours worked per week. After adjusting for other factors, for each additional hour worked per week, the odds of earning over $50,000 was 1.03 times the odds for someone who works one hour less.

- Model Comparison:
  - Using the AIC values from a likelihood ratio test, the second model that included the `hours-per-week` variable had a substantially lower AIC value of 19,678 compared to an AIC of 19,984, indicating a better fit. The lower BIC value from the second model also provides a better fit as it penalizes complexity and overfitting.

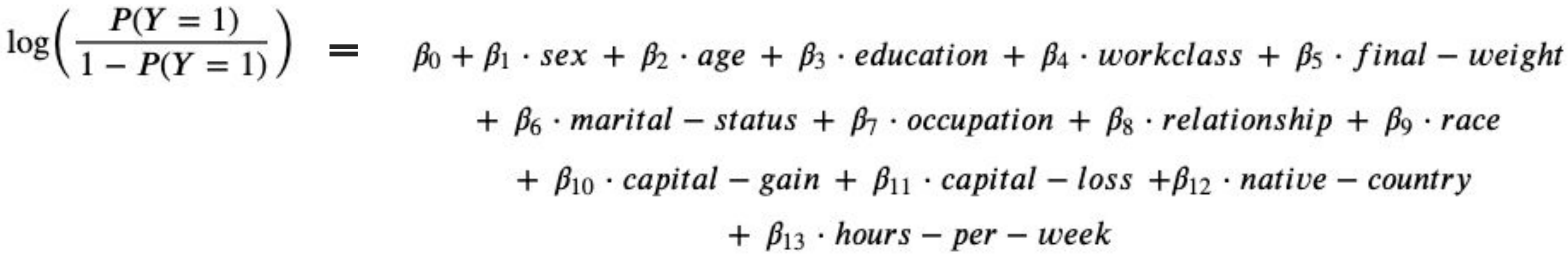
Model	AIC	BIC
Model 1: Excluded `hours-per-week`	19984	20774
Model 2: Included `hours-per-week`	19678	20476

Results

Earnings and Hours worked

Selected Model 2

With Y as the binary response variable such that Y = 1 if the individual earns more than $50,000 or Y = 0 if the individual earns less than $50,000, the model can be expressed as:



$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot age + \beta_3 \cdot education + \beta_4 \cdot workclass + \beta_5 \cdot final - weight$$
$$+ \beta_6 \cdot marital - status + \beta_7 \cdot occupation + \beta_8 \cdot relationship + \beta_9 \cdot race$$
$$+ \beta_{10} \cdot capital - gain + \beta_{11} \cdot capital - loss + \beta_{12} \cdot native - country$$
$$+ \beta_{13} \cdot hours - per - week$$

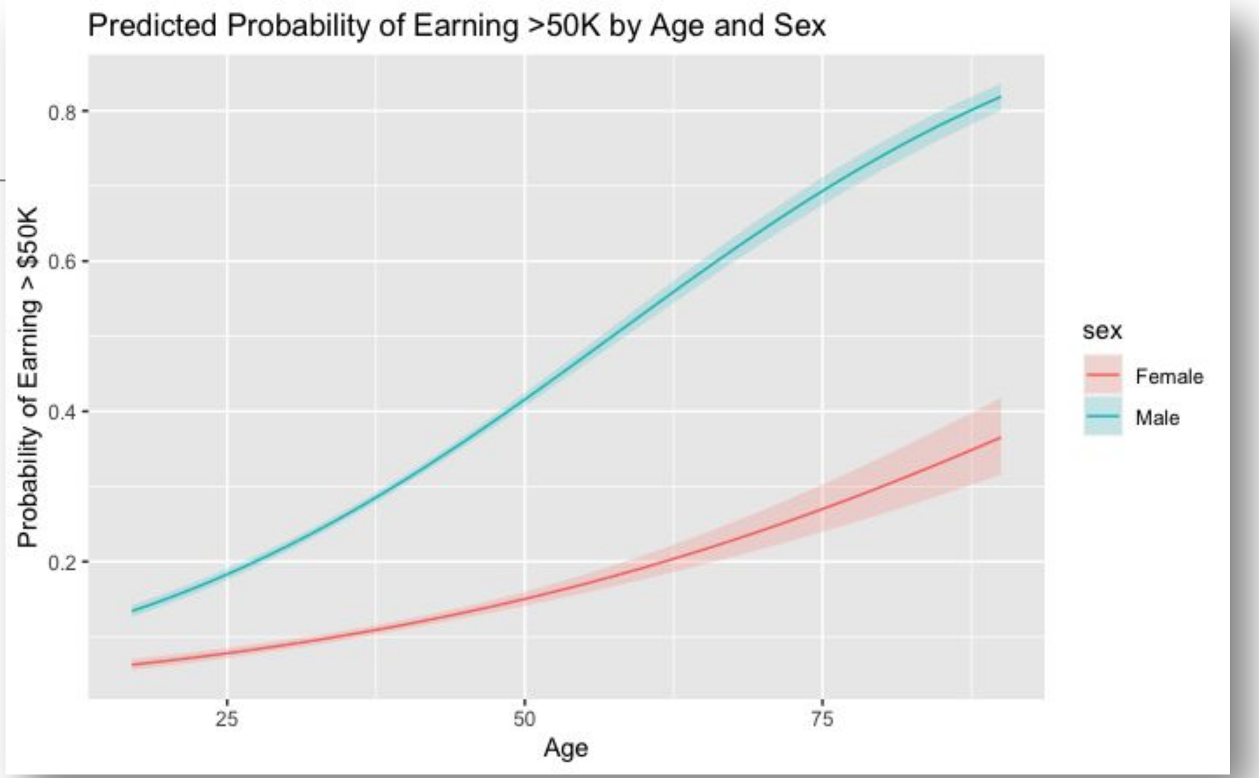
Results

Earnings Between Men and Women

- Males had an estimated probability of 31% of earning more than $50,000, females had an estimated probability of 11%.

- Model 1: Logistic regression model with age and sex.
    - Both predictor variables had a significant effect on the probability of an individual earning more than $50,000.

- Model 2: Logistic regression model with age, sex, and interaction term between age and sex:
    - The addition of the interaction variable 'sex*age' added a strong effect (p-value < 0.001) to the model with a coefficient of 0.016739,translating to an odds ratio of 1.02.
    - Interpretation: For males, each additional year of age increases the odds of earning more than $50K by approximately 1.02 times the odds more than it does for females.



The plot highlights the disparity between genders. It helps visually reinforce the inclusion of the interaction term between sex and age. Both females and males show an increasing income with age – males is more pronounced through all ages.

Results

Earnings Between Men and Women

Model Comparison

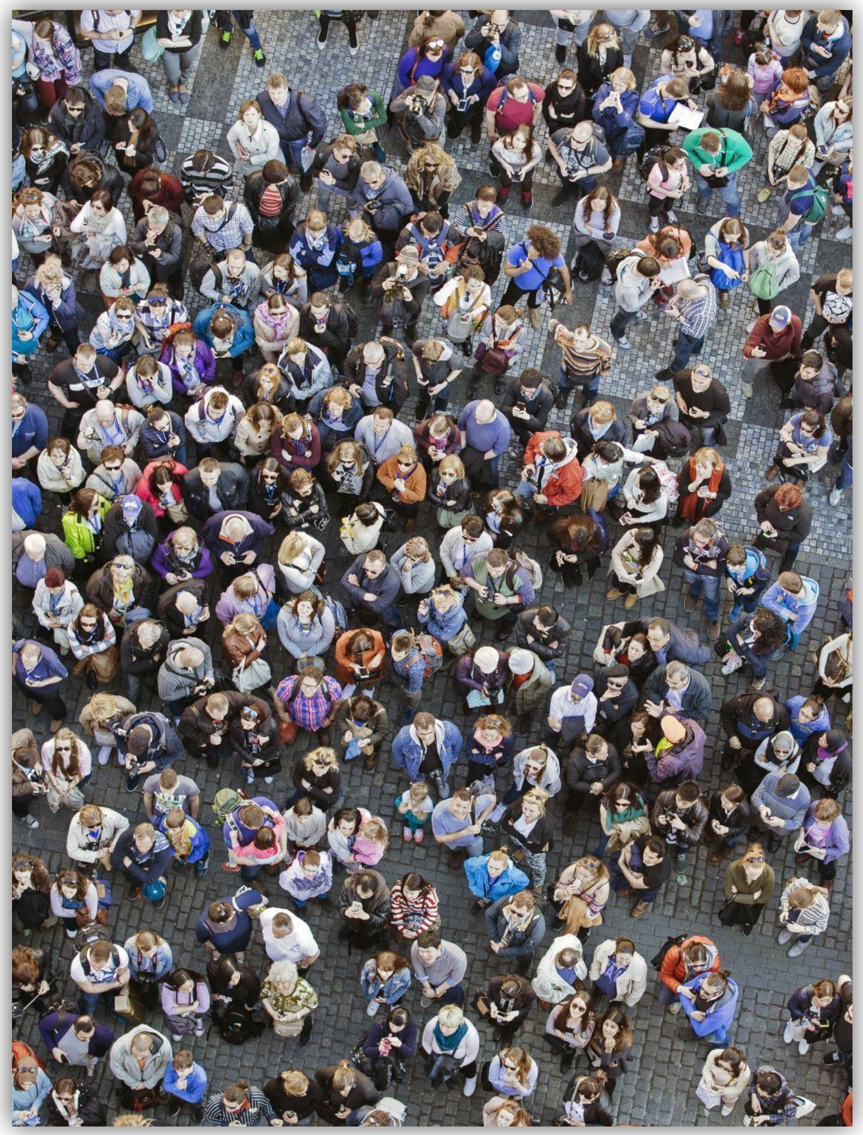
Model	AIC	BIC
Model 1: Excluded Interaction Term	31256	31281
Model 2: Included Interaction Term	31215	31248

Compared two nested logistic regression models using the Akaike Information Criterion (AIC). The second model that included the interaction term had a substantially lower AIC value of 31215 compared to an AIC of 31256, indicating a better fit. The lower BIC value from the second model also provides a better fit as it penalizes complexity and overfitting.

Selected Model 2

With Y as the binary response variable such that Y = 1 if the individual earns more than $50,000 or Y = 0 if the individual earns less than $50,000, the model can be expressed as:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 \cdot \mathbf{sexMale} + \beta_2 \cdot \mathbf{age} + \beta_3 \cdot (\mathbf{sexMale} \cdot \mathbf{age})$$



Scope of Inference

- The data includes a **non-random** sample of working U.S. adults aged 16+ from 1994.
- Excluding groups (e.g., students, retirees, non-workers) is not a representation of the population.
- Findings apply only to individuals similar in demographic and employment characteristics to those in the dataset.

Limitations

- The observational nature of the data allows for identifying associations, not causation.
- Outdated data does not reflect current economy and demographics, work roles, or education impacts.

Discussion/ Conclusion

The use of logistic regression was explored in this analysis to determine if the number of hours that an individual worked were associated with the likelihood of earning more than $50,000 annually. It was confirmed that the relationship between hours worked and earnings of over $50,000 had a strong effect on the performance of the logistic regression model. Additionally, gender disparities were also discovered. Men were substantially more likely to earn over $50,000 than women even after adjusting for age.

It was demonstrated that demographic and employment factors, particularly hours worked and gender, play a significant role in predicting income levels. The use of logistic regression offers valuable insights while highlighting areas for further statistical and policy-oriented investigation.

NEXT STEPS:

- Explore further the gender gap across other variables besides age. Does that disparity between genders exist after accounting for other factors?

- Refit model on a training set, consider using only the most predictive variables for prediction on the test set.