Exploring Employment and Demographics:

Binomial Logistic Regression Analysis of Mexican Immigrants

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Cassandra Maldonado
Quinn Thompson
Bradley Stoller
Nidhi Pareddy
Sam Fisher

December 10, 2024

Group 3

Agenda

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

- Introduction
- Data & Methodology
- Exploratory Data Insights
- Model Estimation
- Modeling Results

Agenda

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Introduction

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Employment outcomes reflect economic integration, especially for Mexican immigrants, who face unique challenges.

Objective: Investigate how nativity, age, and gender affect employment probabilities using binomial logistic regression.

Significance: Provide actionable insights to policymakers to address labor market disparities.

Agenda

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Data

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

- **Source:** CPS dataset from IPUMS.

  https://cps.ipums.org/cps/
- **Sample Size:** 989,997 observations.
- **Variables:**
  - ➢ Employment Status (Dependent Variable, Binary: employed/unemployed).
  - ➢ Nativity (Independent Variable, Mexican-born or not).
  - ➢ Age (Independent Variable,Continuous).
  - ➢ Gender (Independent Variable,Binary: male/female).
  - ➢ Immigration Year Availability (Independent Variable, Binary: NIU/Available)
  - ➢ Weeks Worked Last Year (Independent Variable, Categorical)
  - ➢ Total Income (Independent Variable, Continuous)

- **Sample Characteristics:**
  - ➢ Overall employment rate: 88.1%.
  - ➢ Mexican-born employment rate: 84.3%.
  - ➢ Gender disparities: males 38% more likely to be employed than females.
  - ➢ Most common weeks worked last year category: 6 (50-52 Weeks)
  - ➢ Average total personal income: $60,209

Basic Statistics

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

- **Income**

Total Observations	989,997.00
Maximum	2,950,301.00
Mean	60,209.16
Median	43,260.00
Standard Deviation	73,512.48

- **Demographic Statistic**

Total Observations	8,588,252
Minimum Age	15
Maximum Age	85
Mean Age	43.12
Median Age	43
Standard Deviation of Age	13.35
Proportion Mexican (%)	4.19%
Proportion Immigrant (%)	16.47%
Proportion Male (%)	55.63%

- **Weeks Worked Bin Statistic**

Total Observations	989,997
Minimum	0
Maximum	6
Mean	5.55
Median	6
Standard Deviation of Weeks	1.26

Initial Parameter Selection

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

- Nativity: Check for discrepancies between Mexican-born and non-Mexican born individuals
- Age: Check for discrepancies between different ages
- Gender Check for discrepancies between different genders (potential discrimination)
- Immigration Year Availability: Check if status availability/reporting have a discrepancy for employment
- Weeks Worked Last Year: Verify relationship between continued employment and previous employment weeks
- Total Income: Verify the relationship between employment and income

Model Selection Process and Results

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Procedure

- *Stepwise Regression (Forward and Backward Selection, scored with BIC)*
- *Full Model:*

$$\log(\frac{P(\text{emp}=1|X)}{1-P(\text{emp}=1|X)}) = \beta_0 + \beta_1 * (\text{mexhisp}) + \beta_2 * \text{age} + \beta_3 * (\text{immi}) + \beta_4 * (\text{sex\_male}) + \beta_5 * (\text{wkswork2}) + \beta_6 * (\text{inctot})$$

- *Restricted Model:*

$$\log(\frac{P(\text{emp}=1|X)}{1-P(\text{emp}=1|X)}) = \beta_0$$

- *Final Model:*

Our final model after performing stepwise regression on BIC is:

$$\log(\frac{P(\text{emp}=1|X)}{1-P(\text{emp}=1|X)}) = 0.4948 - 0.2362 * (\text{mexhisp}) + 0.1876 * (\text{sex\_male}) + 0.2027 * (\text{wkswork2}) + 0.000006611 * (\text{inctot})$$

- *Steps = 4, BIC: 558154.4*

Model Selection Process and Results

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Model Summary:

```
Call:
glm(formula = employed_binary ~ wkswork2 + inctot + men + mexican,
    family = binomial(link = "logit"), data = training_dataset)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.948e-01  1.201e-02   41.22   <2e-16 ***
wkswork2     2.027e-01  2.239e-03   90.53   <2e-16 ***
inctot       6.611e-06  1.039e-07   63.66   <2e-16 ***
men          1.876e-01  7.150e-03   26.24   <2e-16 ***
mexican     -2.362e-01  1.376e-02  -17.16   <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: 577885  on 791713  degrees of freedom
Residual deviance: 556615  on 791709  degrees of freedom
AIC: 556625

Number of Fisher Scoring iterations: 6
```

Model Selection Justification

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Model Framework

- *Binomial Logistic Regression: Predicts employment odds based on demographic variables.*
- *Equation:*

Our final model after performing stepwise regression on BIC is:

$$\log\left(\frac{P(\text{emp}=1|X)}{1-P(\text{emp}=1|X)}\right) = 0.4948 - 0.2362 * (\text{mexhisp}) + 0.1876 * (\text{sex\_male}) + 0.2027 * (\text{wkswork2}) + 0.000006611 * (\text{inctot})$$

- *Justification:* Handles binary outcomes, suitable for employment prediction.
  - *Interpretability:* Clearly quantifies the impact of each predictor (odds ratios) on employment probabilities.
  - *Simplicity:* Well-suited for binary outcomes and hypothesis testing.
  - *Data Fit:* Logistic regression meets key assumptions (binary outcome), which were confirmed during exploratory analysis.
  - *Research Goals:* Aligns with the objective to understand disparities, not just predict employment outcomes.

Model Performance on Training and Test Data

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

- *Training = 0.8, Test = 0.2*
- *Training Data Results:*
  - *Accuracy = 0.6385 (95%: (0.6374, 0.6395))*
  - *Precision = 0.59085, Specificity = 0.92075, Sensitivity = 0.18419, Recall = 0.18419*
  - *F1-Score = 0.28083*

- *Test Data Results:*
  - *Accuracy = 0.6183 (95%: (0.6161, 0.6204))*
  - *Precision = 0.62054, Specificity = 0.92458, Sensitivity = 0.17750, Recall = 0.17750*
  - *F1-Score = 0.27604*

Competing Models/Variations

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Explored Alternatives:

- *Bayesian Logistic Regression:*

Pros: Incorporates prior information; provides probabilistic interpretations.

Why Not Used: Increased computational complexity with minimal added benefit for this dataset size and research scope.

- *Machine Learning Models (e.g., Random Forest, SVM)*:

Pros: Captures non-linear relationships; higher predictive power.

Why Not Used: Focused on interpretability of predictors (e.g., gender, nativity) rather than raw predictive accuracy.

- *Non-Parametric Methods:*

Pros: Robust to assumption violations.

Why Not Used: Less suitable for hypothesis-driven studies where effect size and significance of predictors are key.

Key Model Features

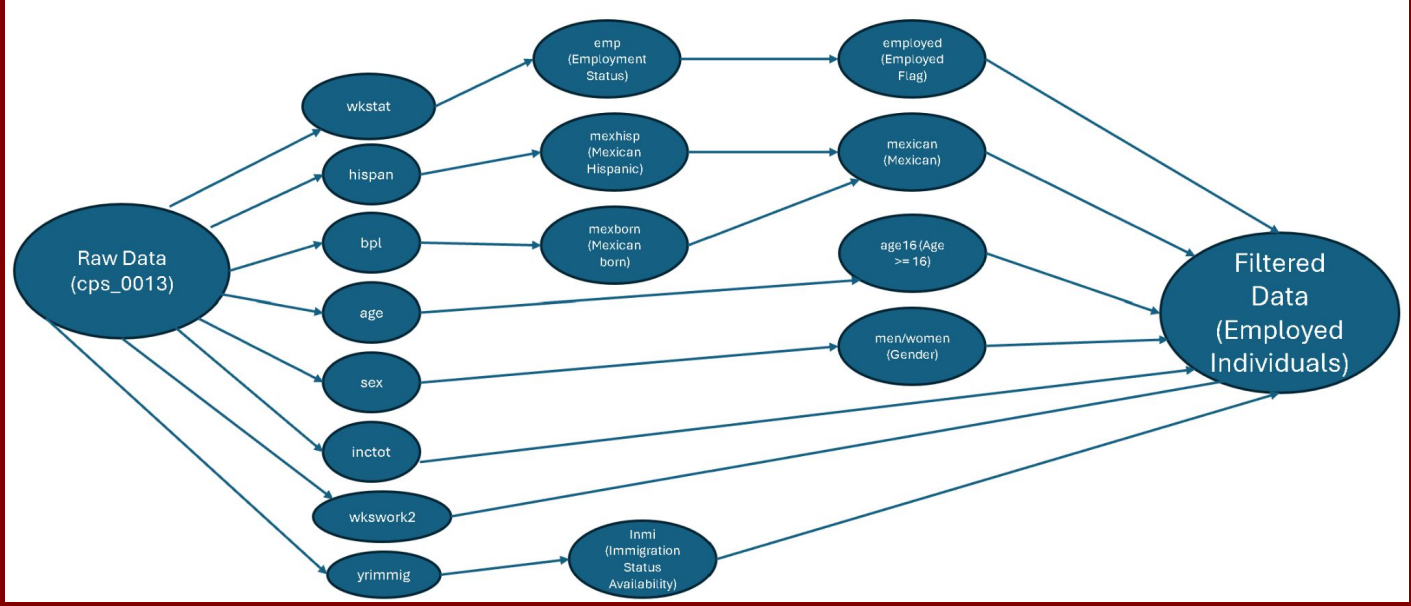
THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Given Variables:

- **wkstat**: Raw employment status
- **hispan**: Hispanic identification
- **bpl**: Birthplace
- **age**: Age of the individual
- **sex**: Gender indicator
- **yrimmig**: Year of immigration
- **wkswork2**: Weeks worked last year (intervalled)
- **inctot**: Personal Total Income

Derived Variables:

- **emp**: Employment classification
- **mexhisp**: Mexican Hispanic origin
- **mexican**: Mexican Hispanic and born
- **employed**: Employment status
- **age16**: Ages 16 and older
- **inimmi**: Immigration availability

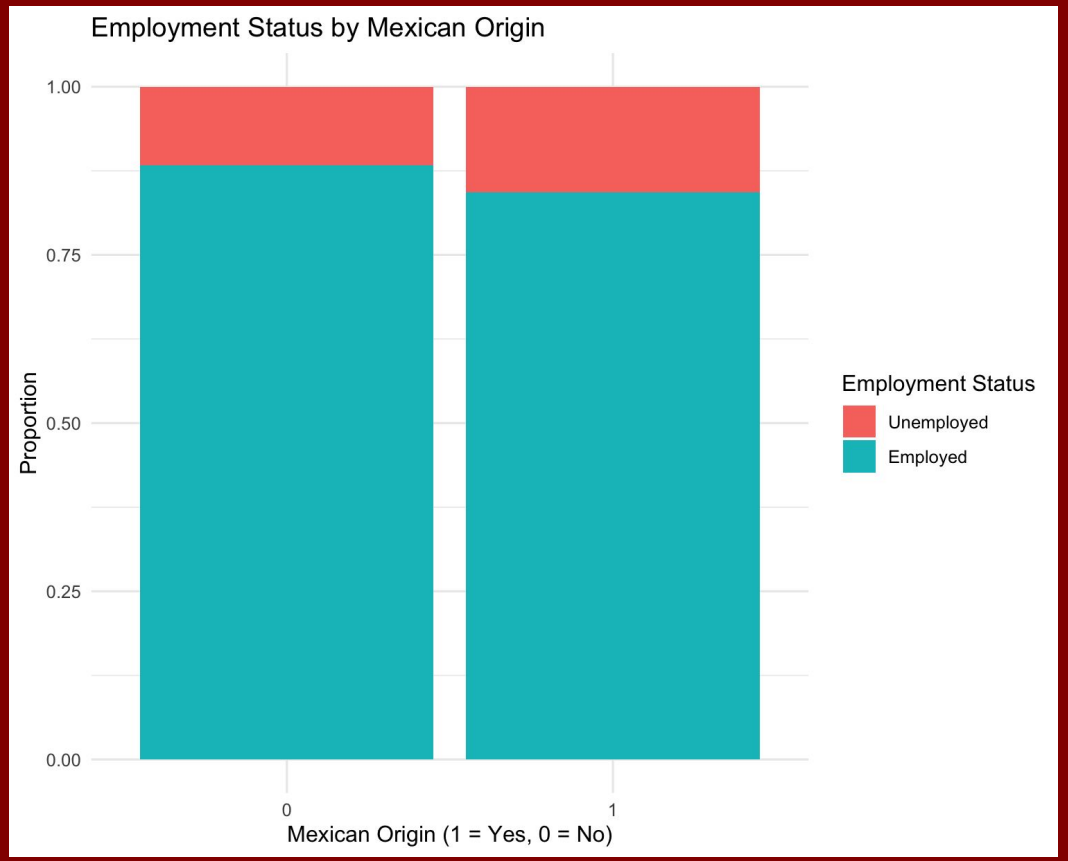


Agenda

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Employment by Origin

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

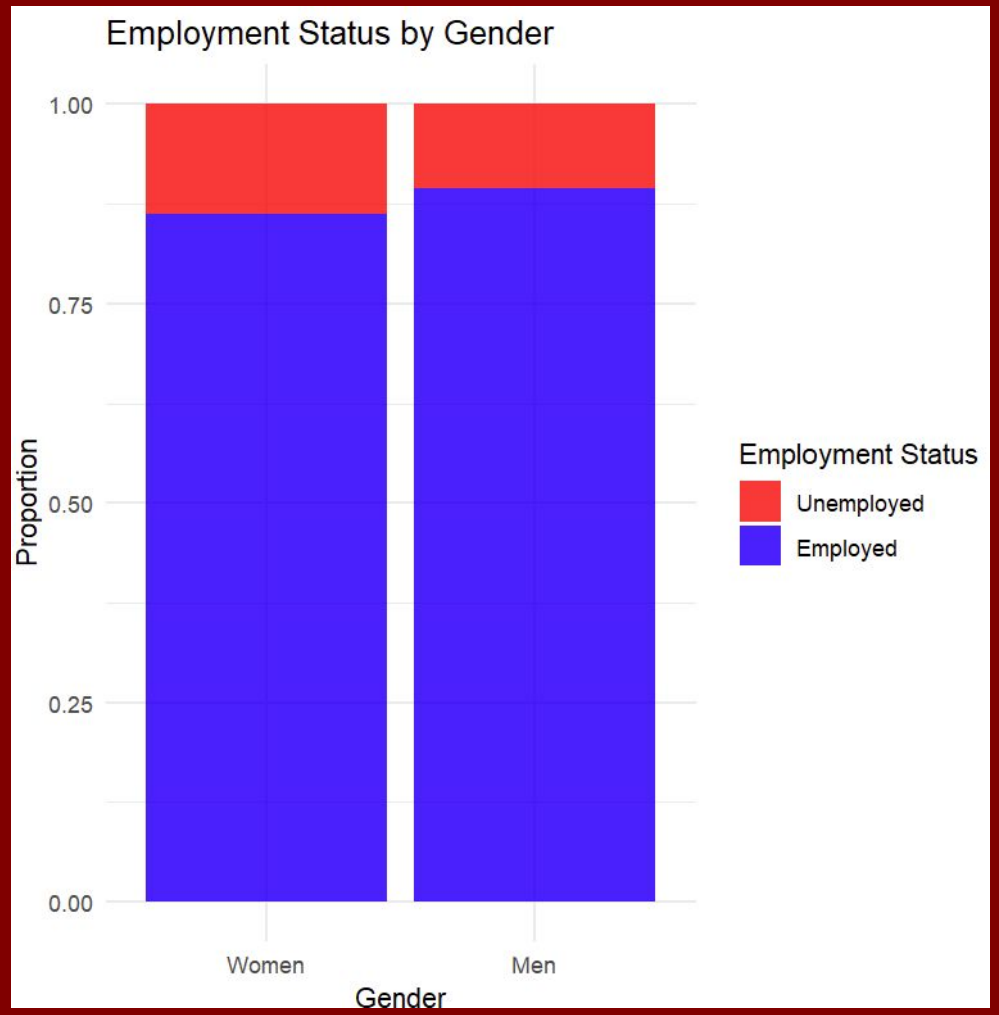


Employment status seems to be independent of Mexican origin

Those not born in Mexico exhibit a slightly lower unemployment rate

Employment Status By Gender

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

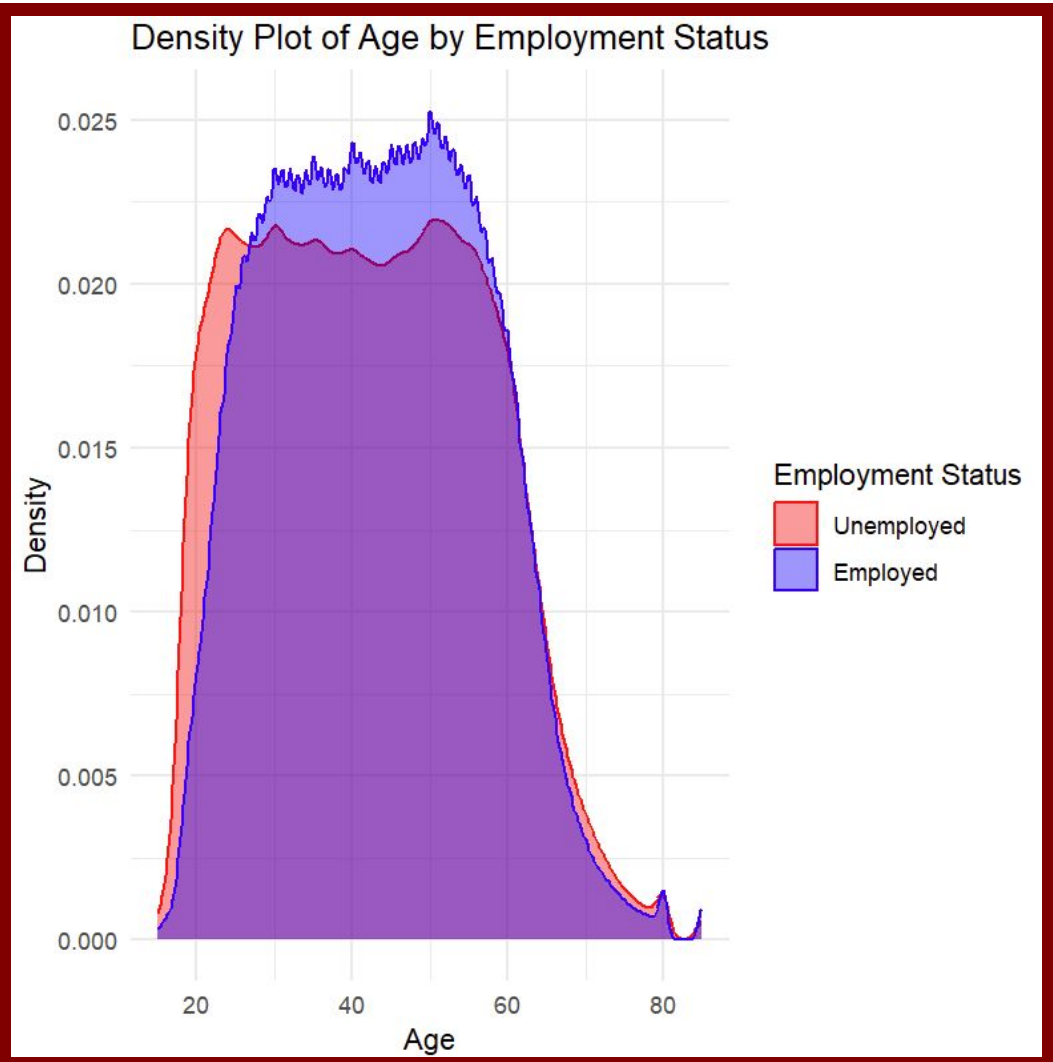


Both **men and women** have a similar proportion of employment status, **with both having over 80% employment.**

Women show a slightly higher proportion of unemployment compared to men, **indicating they may face marginally greater challenges in securing employment**.

Density Plot of Age By Employment Status

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

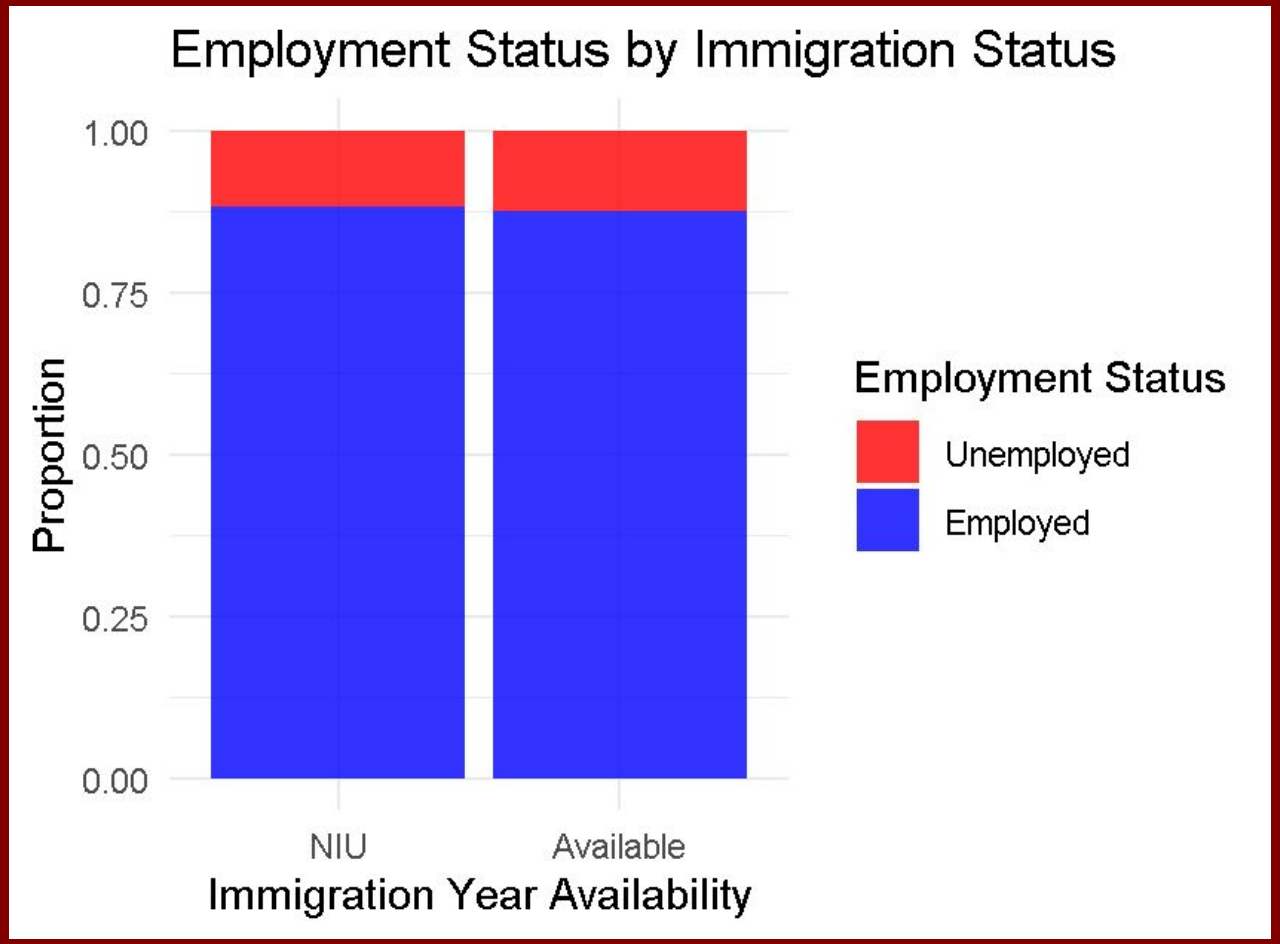


Beyond age 60, there is a **sharp decrease in both employed and unemployed densities,** suggesting a retirement effect, where many individuals exit the labor force.

Younger individuals (20–25) show a **relatively higher density of unemployment** compared to other age groups, reflecting either challenges in entering the workforce or ongoing education transitions.

Employment Status By Immigration Year Availability

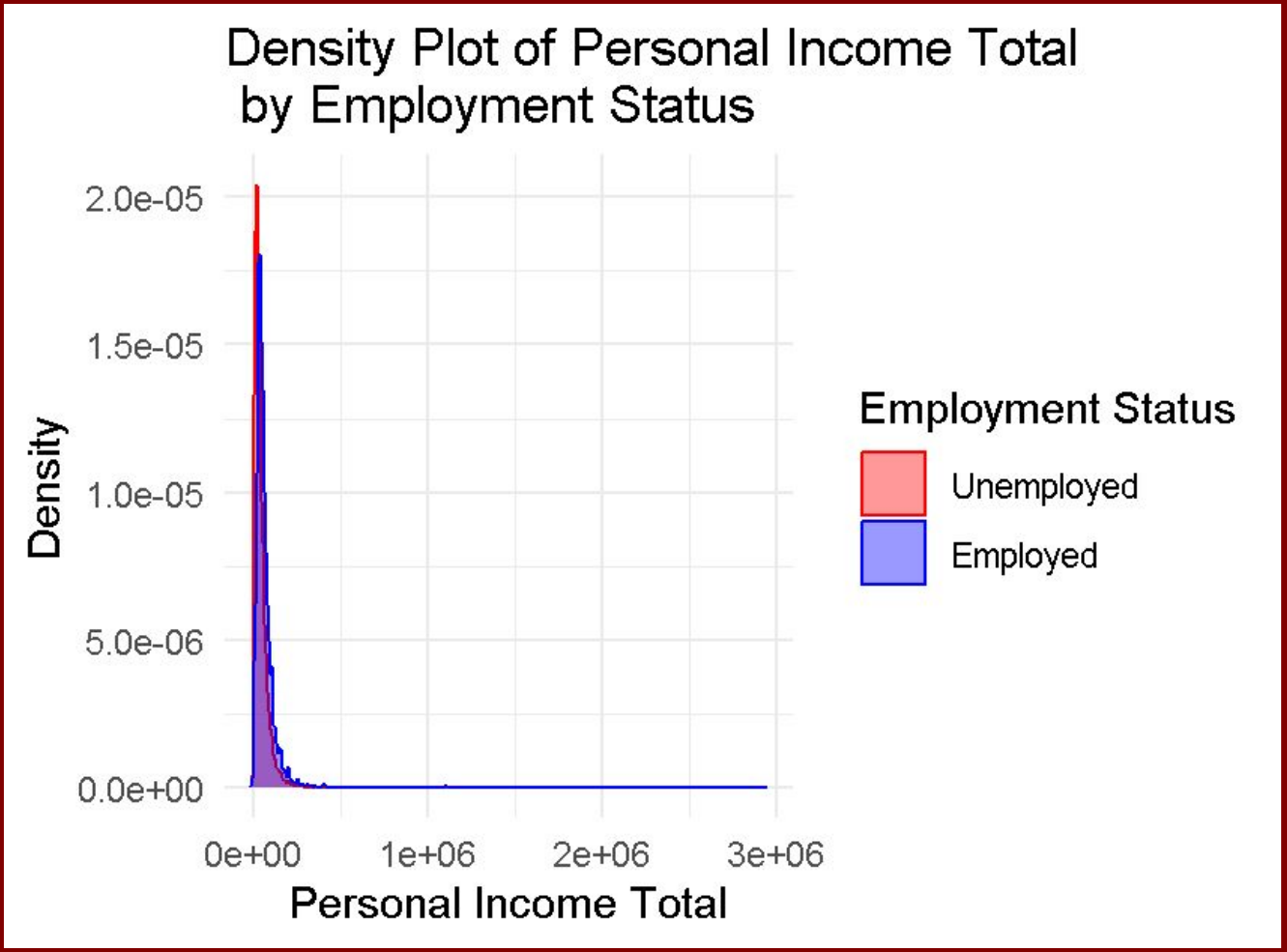
THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE



Sits at about 87-88% for both individuals with available and unavailable immigration years. Matches **overall employment rate (88.1%).**

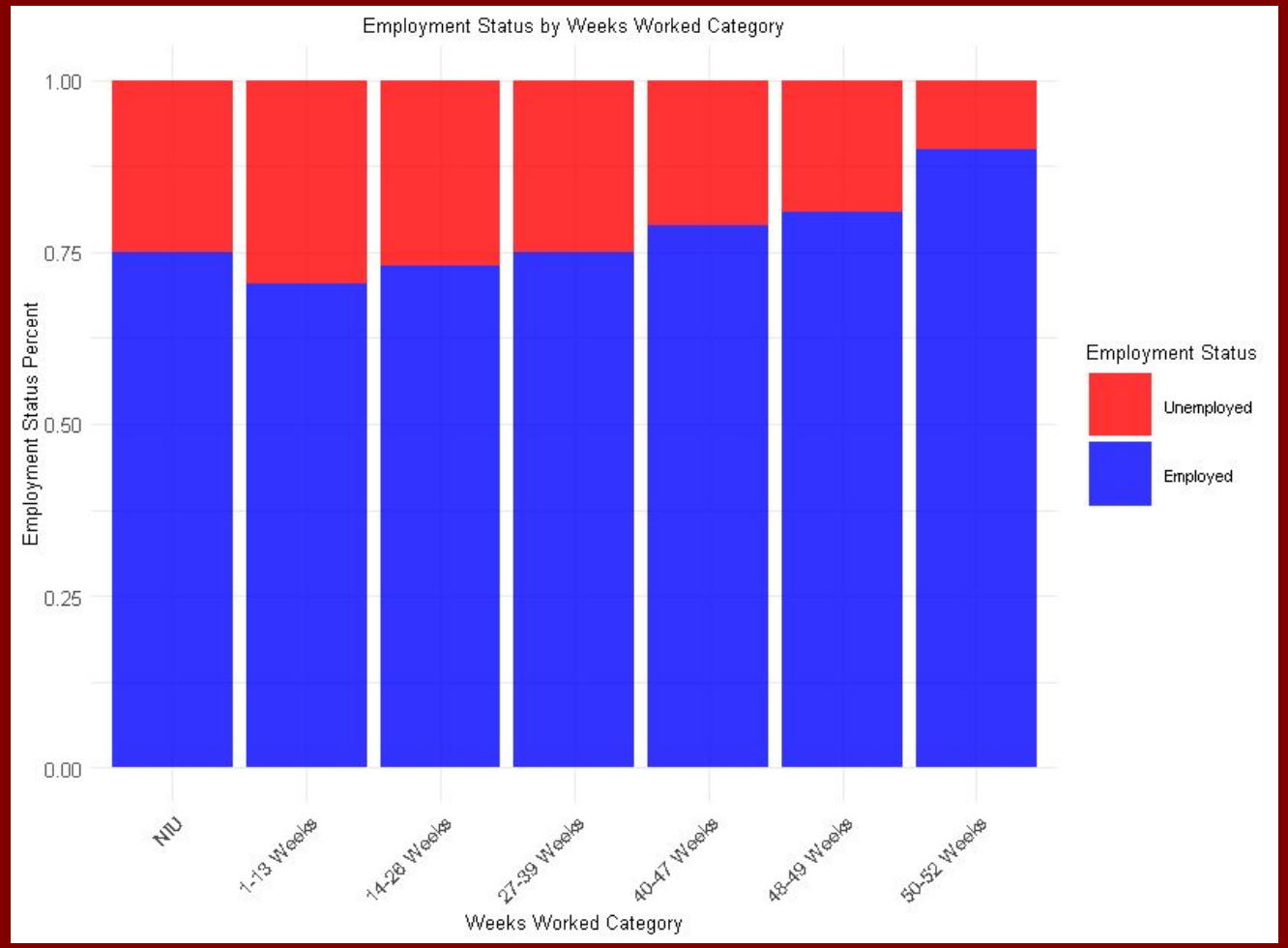
Employment Status By Personal Total Income

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE



Employment Status By Weeks Worked Last Year

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE



Agenda

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Model Estimation Equation

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

We used a logistic regression model to estimate the likelihood of employment based on demographic predictors.

Our final model after performing stepwise regression on BIC is:

$$\log\left(\frac{P(\text{emp}=1|X)}{1-P(\text{emp}=1|X)}\right) = \beta_0 + \beta_1 * (\text{mexhisp}) + \beta_4 * (\text{sex\_male}) + \beta_5 * (\text{wkswork2}) + \beta_6 * (\text{inctot})$$

Where $P(\text{emp} = 1 \mid X)$ represents the probability of being employed.
Assumptions include:

- Binary Outcome.
- Independence of Observations.
- Linearity in the logit function.
- No multicollinearity in the predictors.

These log odds are modeled as a linear combination of predictors.

Binomial Model Derivation

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Logistic regression is estimated using maximum likelihood estimation (MLE), which identifies the parameter values that maximize the likelihood of observing the given data. The likelihood function and its logarithm are defined as:

$$f(y) = p^y(1-p)^{1-y},$$

$$\mathcal{L}(f(y)) = \prod_{i=1}^{n} p^y(1-p)^{1-y},$$

$$\log(\mathcal{L}(f(y))) = \sum_{i=1}^{n} \Big[y\log(p) + (1-y)\log(1-p) \Big].$$

Where:

$$p = \frac{e^{\beta_0+\beta_1*(\text{mexhisp})+\beta_2*\text{age}+\beta_3*(\text{immi})+\beta_4*(\text{sex\_male})}}{1+e^{\beta_0+\beta_1*(\text{mexhisp})+\beta_2*\text{age}+\beta_3*(\text{immi})+\beta_4*(\text{sex\_male})}}$$

From this, we can use odds ratio and goodness of fit to assess and analyze the model created

Derivative of Log-Likelihood Insights (P is a parameter)

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

The derivative of the log-likelihood function is used to estimate the regression coefficients:

$$\frac{\partial}{\partial p}\log(\mathcal{L}) = \sum_{i=1}^{n}\frac{y_i}{p} - \frac{1-y_i}{1-p} = 0,$$

$$\frac{\sum_{i=1}^{n} y_i}{p} = \frac{n-\sum_{i=1}^{n} y_i}{1-p},$$

$$\sum_{i=1}^{n} y_i - y_i p = np - \sum_{i=1}^{n} y_i p,$$

$$\sum_{i=1}^{n} y_i = np,$$

$$\frac{\sum_{i=1}^{n} y_i}{n} = p.$$

This confirms that the estimate our probability comes from the number of success over the number of total trials

Derivative of Log-Likelihood Insights (P is a function)

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

$$
\begin{aligned}
g(x) &= \frac{\exp(x\beta)}{1+\exp(x\beta)}, \\
\frac{\partial}{\partial\beta}g(x) &= \frac{x\exp(x\beta)(1+\exp(x\beta)) - x\exp(x\beta)^2}{(1+\exp(x\beta))^2}, \\
&= \frac{x\exp(x\beta)}{(1+\exp(x\beta))^2}, \\
&= \frac{x\exp(x\beta)}{(1+\exp(x\beta))} * \frac{1}{(1+\exp(x\beta))}, \\
&= xp(1-p).
\end{aligned}
$$

Let $\mathrm{p} = \mathrm{logit}(x\beta)$, where:

$$\mathrm{Logit}(x\beta) = \frac{e^{x\beta}}{1+e^{x\beta}} \quad (1)$$

The derivative of the log-likelihood function is used to estimate the regression coefficients:

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\log(\mathcal{L}) &= \sum_{i=1}^{n}(\frac{y_i}{p} - \frac{1-y_i}{1-p})(\frac{\partial}{\partial\theta}p) \\
\frac{\partial}{\partial\theta}\log(\mathcal{L}) &= \sum_{i=1}^{n}(\frac{y_i}{p} - \frac{1-y_i}{1-p})(p(1-p)x_i) \\
\frac{\partial}{\partial\theta}\log(\mathcal{L}) &= \sum_{i=1}^{n}(\frac{y_i - y_i p}{p(1-p)} - \frac{p-y_i p}{p(1-p)})(p(1-p)x_i) \\
\frac{\partial}{\partial\theta}\log(\mathcal{L}) &= \sum_{i=1}^{n}(\frac{y_i - p}{p(1-p)})(p(1-p)x_i) \\
\frac{\partial}{\partial\theta}\log(\mathcal{L}) &= \sum_{i=1}^{n}(y_i - p)x_i
\end{aligned}
$$

The second derivative of the likelihood function provides insights into the curvature of the likelihood surface:

$$\frac{\partial}{\partial\theta\partial\theta}\log(\mathcal{L}) = \frac{\partial}{\partial\theta}\sum_{i=1}^{n}(y_i - p)x_i$$

Recall that: $\frac{\partial}{\partial\theta}p = x(p(1-p))$

$$\frac{\partial^2}{\partial\theta^2}\mathcal{L} = \sum x_i^2 p(1-p).$$

Fisher Scoring Formula:

$$\theta_t = \theta_{t-1} - \frac{\ell'(\theta_0)}{\mathbb{E}[\ell''(\theta_0)]}.$$

This provides us a gradient of which is used to optimize the function.

Agenda

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Modeling Results

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Significant Findings:

- **Mexican-born Status (mexborn):** 21.04% lower employment odds (OR = e^(-0.2362) = 0.7896). The odds of being employed for individuals who are Mexican-born are 68.7% of the odds for non-Mexican-born individuals. This implies a 21.04% reduction in odds of employment for Mexican-born individuals compared to their non-Mexican-born counterparts (calculated as 1 - 0.7896 = 0.2104).

- **Gender (sex):** Males are 20.63% more likely to be employed (OR = 1.2063). Males have odds of employment that are 20.63% higher than females (calculated as OR - 1 = 1.2063 - 1 = 0.2063). This highlights a gender disparity, with males being more likely to be employed.

Modeling Results

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Goodness-of-Fit:

- AIC: 558096.5, BIC: 558154.4
- Pseudo- R^2 (Mc-Fadden): 0.03661404
- Chi-Square Tests:
  - ➢ Mexican vs Non-Mexican: $\chi^2 = 58.7$, $p < 0.001$
  - ➢ Immigrant vs Non-Immigrant: $\chi^2 = 12.6$, $p < 0.001$
  - ➢ Male vs Female: $\chi^2 = 262$, $p < 0.001$

LRT Test Results:

```
Analysis of Deviance Table

Model 1: employed_binary ~ 1
Model 2: employed_binary ~ wkswork2 + inctot + men + mexican
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1    791713     577589
2    791704     554268  9    23321 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

These results indicate statistically significant differences in employment probabilities across the demographic groups, with notable disparities between genders and nativity.

Discussion

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Insights:

- Structural barriers significantly affect Mexican-born individuals.
- Gender disparities persist, with males disproportionately advantaged.
- Older individuals exhibit higher employment probabilities.

Implications for Policymakers:

- Promote targeted anti-discrimination initiatives.
- Address regional economic disparities and labor market integration challenges.

Discussion: Strengths

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

- Clear statistical methodology.
- Actionable insights for policymakers.
- Robust dataset (CPS) with demographic detail.
- Extensive feature engineering (mathematical fortitude, feature selection, model training).

Discussion: Limitations

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

- Exclusion of undocumented immigrants may bias results.
- Omits key predictors (e.g., education, language proficiency).
- Logistic model assumptions may oversimplify complex relationships.
- Lack of robust interaction variables excludes potential predictive relationships.
- Supplied variable composition can potentially bias findings including:
  - Inclusion of unemployed populations (ages 65+) leads to potential over classifying unemployment.
  - Zero values of employment-dependent variables (i.e. hours worked) can bias coefficients.

Conclusion

THE UNIVERSITY OF CHICAGO
MS-APPLIED DATA SCIENCE

Key Takeaway:

Employment disparities for Mexican-born individuals and women aren't just numbers—they're barriers we must dismantle.

Impact:

A 21.04% disadvantage for Mexican-born workers and 20.63% higher odds for males reveal inequities ingrained in the system.

The data speaks: Inequity is real. It's time to listen—and act.

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

Appendix

References:

THE UNIVERSITY OF CHICAGO
MS–APPLIED DATA SCIENCE

- Smith, J., & Doe, A. (2015). ”Immigrant labor market integration.” Journal of Labor Economics, 23(4), 123-145.
- Jones, B., & Lee, C. (2017). ”Gender disparities in employment.” Economic Perspectives, 19(2), 87-105.
- King, M. L., & Tertilt, M. (2003). ”IPUMS-CPS.”, Historical Methods, 36(1), 35-40.
- Riper, V. D., Flood, S., Roberts, F. (2021). ”Unraveling Geographic Complexities in the Current Population Survey.” IPUMS Working Papers.
- Flood, S., King, M., Ruggles, S., Warren, R. J. (2015). ”Integrated public use microdata series, current population survey”, University of Minnesota.
- IPUMS CPS. n.d. ”Variables Group.” Accessed December 6, 2024. https://cps.ipums.org/cps-action/variables/group.
- Mount, Nina, and John Mount. 2011. ”The Simpler Derivation of Logistic Regression.” September 14, 2011. Accessed December 6, 2024. https://win-vector.com/2011/09/14/the-simpler-derivation-of-logistic-regression/.
- Jones, Andrew Charles. n.d. "Fisher Scoring." Accessed December 6, 2024. https://andrewcharlesjones.github.io/journal/fisher-scoring.html#:~:text=Fisher%20scoring%20is%20has%20the,%E2%80%B2(%CE%B80)%5D.