Lab3

Ben Tankus

4/16/2021

Questions

The data set `ResumeNames` in the *AER* package contains information about 4870 fictitious resumes, sent to real employers as part of an experiment about racial discrimination in hiring. The binary response for each resume is whether the employer called back. Although the original research question was about racial discrimination, there are many kinds of questions you might answer using these data.

Consider: - Restricting your analysis to an interesting subset of the data - Transforming and combining input variables - Exploring interactions

Some example questions (which you may use or modify if you want)

- (How) does the relationship between callback probability and resume quality differ by applicant race?
- How many additional years of experience is having a white name vs a black name "worth" in terms of callback probability?
- Does callback probability differ between applicants who meet stated job requirements and applicants who don't? For instance, you might compare applicants who do or not meet the stated minimum number of years of experience, or applicants who do or do not list computer skills when applying to jobs for which computer skills are ostensibly required.
- Does callback probability differ between male and female applicants by industry or position?

Be creative!

3. Use a logisitic regression model to address the question you posed in 2. Be sure to examine the fit of your model, and write a few sentences about your interpretation of the model as it addresses the question you posed.
4. Install and load the *AER* package, and read the help file for the `ResumeNames` data.

```
#install.packages('AER')
library('AER')
```

```
## Warning: package 'AER' was built under R version 4.0.5
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.0.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.0.3

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 4.0.4

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Warning: package 'sandwich' was built under R version 4.0.5

## Loading required package: survival
```

```
#?ResumeNames
library(vcdExtra)
```

```
## Warning: package 'vcdExtra' was built under R version 4.0.4

## Loading required package: vcd

## Warning: package 'vcd' was built under R version 4.0.4

## Loading required package: grid

## Loading required package: gnm

## Warning: package 'gnm' was built under R version 4.0.4

##
## Attaching package: 'vcdExtra'

## The following object is masked from 'package:carData':
##
##     Burt
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.3
```

```
## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.3

## Warning: package 'tidyr' was built under R version 4.0.3

## Warning: package 'readr' was built under R version 4.0.3

## Warning: package 'purrr' was built under R version 4.0.3

## Warning: package 'dplyr' was built under R version 4.0.3

## Warning: package 'forcats' was built under R version 4.0.3

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x dplyr::recode()    masks car::recode()
## x purrr::some()      masks car::some()
## x dplyr::summarise() masks vcdExtra::summarise()
```

```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':
##
##     set_names

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(ggplot2)
logistic <- function(x){exp(x)/(1 + exp(x))}
```

2. Come up with a question about the probability of `callback` (the binary response) that can be answered using at least one (but no more than 5) of the 26 available predictor variables.

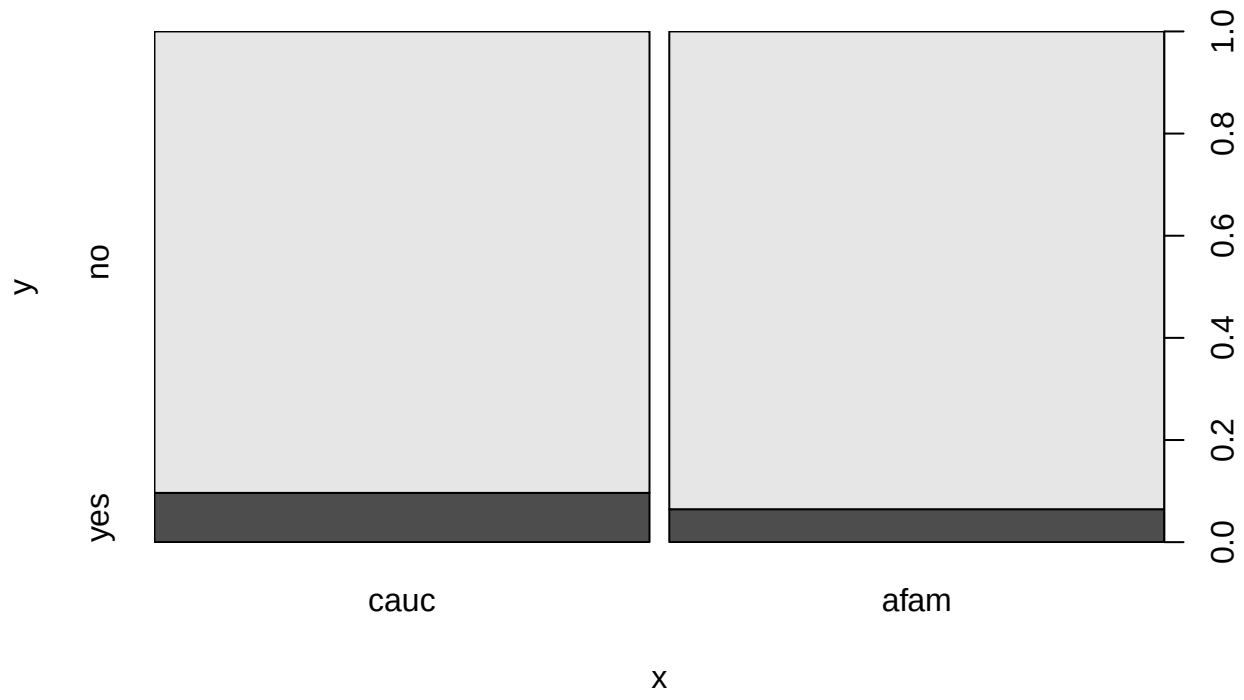
Question: Does ethnicity play a factor in callback?

3. Use a logisitic regression model to address the question you posed in 2. Be sure to examine the fit of your model, and write a few sentences about your interpretation of the model as it addresses the question you posed.

```r
data('ResumeNames')
summary(ResumeNames)
```

```
##       name         gender     ethnicity   quality      call           city
##  Tamika : 256   male  :1124   cauc:2435   low :2424   no :4478   boston :2166
##  Anne   : 242   female:3746   afam:2435   high:2446   yes: 392   chicago:2704
##  Allison: 232
##  Latonya: 230
##  Emily  : 227
##  Latoya : 226
##  (Other):3457
##       jobs         experience     honors     volunteer  military   holes
##  Min.   :1.000   Min.   : 1.000   no :4613   no :2866   no :4397   no :2688
##  1st Qu.:3.000   1st Qu.: 5.000   yes: 257   yes:2004   yes: 473   yes:2182
##  Median :4.000   Median : 6.000
##  Mean   :3.661   Mean   : 7.843
##  3rd Qu.:4.000   3rd Qu.: 9.000
##  Max.   :7.000   Max.   :44.000
##
##  school     email      computer   special    college       minimum
##  no :2145   no :2536   no : 874   no :3269   no :1366   none   :2746
##  yes:2725   yes:2334   yes:3996   yes:1601   yes:3504   some   :1064
##                                                        2      : 356
##                                                        3      : 331
##                                                        5      : 163
##                                                        1      : 142
##                                                        (Other):  68
##  equal                wanted     requirements reqexp     reqcomm    reqeduc
##  no :3452   manager       : 741   no :1036     no :2750   no :4262   no :4350
##  yes:1418   supervisor    : 376   yes:3834     yes:2120   yes: 608   yes: 520
##             secretary     :1621
##             office support: 578
##             retail sales  : 818
##             other         : 736
##
##  reqcomp    reqorg                                industry
##  no :2741   no :4516   manufacturing                    : 404
##  yes:2129   yes: 354   transport/communication          : 148
##                        finance/insurance/real estate    : 414
##                        trade                            :1042
##                        business/personal services       :1304
##                        health/education/social services: 754
##                        unknown                          : 804
```

```r
# Viz
plot(ResumeNames$ethnicity, ResumeNames$call)
```



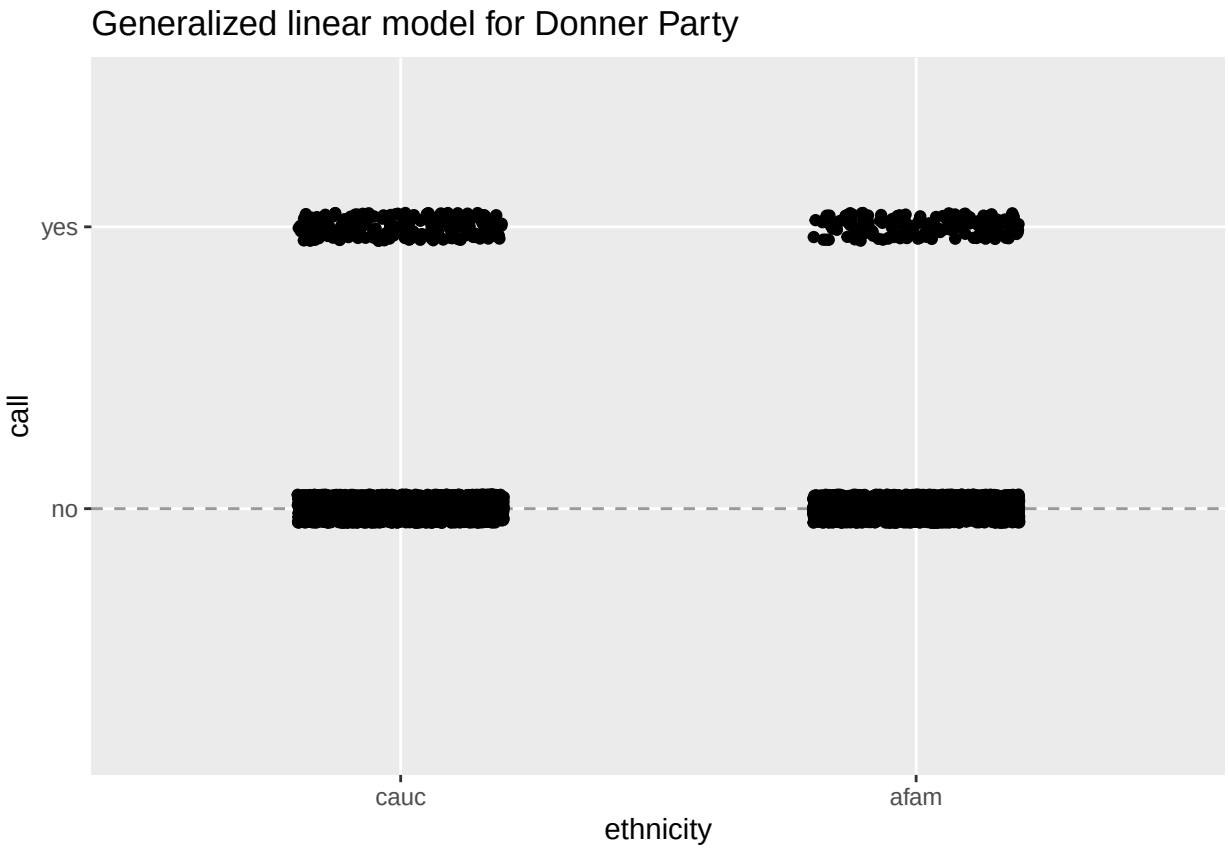

```
glm2 <- glm(call ~ ethnicity, family = binomial(link = "logit"), data = ResumeNames)

summ <- summary(glm2)

summ$deviance
```

```
## [1] 2709.938
```

Answer: Ethnicity does play a factor in callbacks. Black applicants were less likely to get a call back at a 4.45e-05 level of significance on a 95% confidence level. This model doesn't seem to fit well, as the visual does not have a valid fit line.

```
# Obtain 95% pointwise confidence bands from predict.glm()
glm_pred <- predict.glm(glm2, type="link", se.fit=TRUE)
low <- glm_pred$fit - 1.96 * glm_pred$se.fit
upp <- glm_pred$fit + 1.96 * glm_pred$se.fit

# back-transform everything to the data scale
glm_fit <- logistic(glm_pred$fit)
glm_lower <- logistic(low)
```

```r
glm_upper <- logistic(upp)

# augment the Donner data frame
aug_resume <- as.data.frame(cbind(ResumeNames, glm_fit, glm_lower, glm_upper))

# Big plot
ggplot(data = aug_resume) +
  # plot jittered data
  geom_jitter(aes(x = ethnicity,
                  y = call),
                  height = 0.05, width = 0.2) +

  # plot fitted lines
  geom_line(aes(x = ethnicity,
                y = glm_fit)) +

# plot 95% pointwise confidence bands
  geom_ribbon(aes(x = ethnicity,
                  ymin = glm_lower,
                  ymax = glm_upper),
              alpha = 0.2) +

  # plot reference lines at 0 and 1 (minimum and maximum possible probabilities)

  geom_hline(yintercept = 0, lty = 2, alpha = 0.4) +
  geom_hline(yintercept = 1, lty = 2, alpha = 0.4) +
  ggtitle("Generalized linear model for Donner Party")
```



It seems this model does not fit very well as there is no line here. It may fit better with numerical values.