Multinomial (Logistic) Regression

AUTHOR
Exercise answers

Generalized Linear Models

- Allow us to fit models with different types of outcomes: binary, nominal, ordinal, count
- Allow us to **perform inference**, i.e., obtain interpretable estimates, quantify uncertainty of those estimates, conduct hypothesis tests

Multinomial Regression

Define π_j as $P(Y = j)$ for a level j of a nominal categorical variable.

The multinomial model is defined as a set of logistic regression models for each probability π_j compared to a baseline level:

$$log(\frac{\pi_{ij}}{\pi_{i1}}) = \beta_{0j} + \beta_{1j}x_{i1} + \ldots + \beta_{pj}x_{ip}$$

What are all of the indices here (j, p, and i)?

j indexes the level of the nominal categorical outcome variable, p indexes the predictors, and i indexes the subjects

Interpretations

Each coefficient estimate has to be interpreted relative to the baseline outcome level

- If x_1 is continuous, $e^{\beta_{1j}}$ is the multiplicative increase/decrease in the **odds** of $Y = j$ vs $Y = 1$ (baseline) when increasing x_1 by one unit
- If x_1 is categorical, $e^{\beta_{1j}}$ is the odds of $Y = j$ vs $Y = 1$ (baseline) for the group with $x_1 = 1$ compared to the reference group of the covariate

Exercise

You are an analyst at a large technology company. The company recently introduced a new health insurance provider for its employees. At the beginning of the year, the employees had to choose one of three different health plan products from this provider. You have been asked to determine which factors influenced the choice in product.

The `health_insurance` data set consists of the following fields:

- `product`: The choice of product of the individual—A, B or C
- `age`: The age of the individual when they made the choice
- `gender`: The gender of the individual as stated when they made the choice
- `household`: The number of people living with the individual in the same household at the time of the choice
- `position_level`: Position level in the company at the time they made the choice, where 1 is is the lowest and 5 is the highest
- `absent`: The number of days the individual was absent from work in the year prior to the choice

```
library(caret) #for confusion matrix
library(nnet) #for multinomial regression
library(tidyverse)

health_insurance <- read.csv("http://peopleanalytics-regression-book.org/data/health_insurance.csv")
```

1. Create factor variables for `product` and `gender`. Then, run the `levels()` function to verify that "A" is the first/reference level.

   ```
   health_insurance <- health_insurance |>
     mutate(product_fac = factor(product),
            gender_fac = factor(gender))

   levels(health_insurance$product_fac)
   ```

   ```
   [1] "A" "B" "C"
   ```

2. Calculate N and % for each product. Are the levels balanced or imbalanced?

```
health_insurance |>
  count(product_fac) |>
  mutate(prop=n/sum(n))
```

```
  product_fac   n      prop
1           A 495 0.3406745
2           B 459 0.3158981
3           C 499 0.3434274
```

The levels are balanced, with approximately 1/3 of the observations in each category

3. Use the code below to fit a multinomial model regressing `product` on all other variables in the dataset.

```
healthinsurance_mod1 <- nnet::multinom(product ~ age + gender_fac +
                                         household + position_level +
                                         absent,
                                       data = health_insurance)
```

```
# weights:  24 (14 variable)
initial  value 1596.283655
iter  10 value 969.042921
iter  20 value 744.786124
final  value 744.682377
converged
```

```
summary(healthinsurance_mod1)
```

```
Call:
nnet::multinom(formula = product ~ age + gender_fac + household +
    position_level + absent, data = health_insurance)

Coefficients:
  (Intercept)       age gender_facMale gender_facNon-binary  household
B    -4.60100 0.2436645    -2.38259765            0.2523409 -0.9677237
C   -10.22617 0.2698141     0.09670752           -1.2715643  0.2043568
  position_level      absent
B     -0.4153040 0.011676034
C     -0.2135843 0.003263631

Std. Errors:
  (Intercept)       age gender_facMale gender_facNon-binary  household
```

```
B   0.5105532 0.01543139      0.2324262             1.226141 0.06943089
C   0.6197408 0.01567034      0.1954353             2.036273 0.04960655
  position_level     absent
B     0.08916739 0.01298141
C     0.08226087 0.01241814

Residual Deviance: 1489.365
AIC: 1517.365
```

```r
odds_scale_coefs <- exp(summary(healthinsurance_mod1)$coefficients)

data.frame(t(odds_scale_coefs)) #formats coefficients on odds scale so they're easier to see
```

```
                            B            C
(Intercept)          0.01004179 3.621021e-05
age                  1.27591615 1.309721e+00
gender_facMale       0.09231048 1.101538e+00
gender_facNon-binary 1.28703467 2.803927e-01
household            0.37994694 1.226736e+00
position_level       0.66013957 8.076841e-01
absent               1.01174446 1.003269e+00
```

- How many coefficient estimates do we obtain from this model fit?
    - We calculate the number of coefficient estimates in a multinomial model with $(P+1) \times (J-1)$, where P is the number of predictor terms (account for levels of categorical predictors/interaction terms and add one for the intercept), and J is the number of levels of the categorical outcome (subtract one because we don't have estimates for the baseline level). Here, $P = 6$ and $J = 3$, so we get $7 \times 2 = 14$ coefficient estimates.
- What information do we **not** get from the summary output here that we've seen for the models we've fit before?
    - We only see estimates and standard errors here, so we don't get t-values and p-values
- Write interpretations for the coefficient estimates on the odds scale for `gender=male`, `age`, and `household` for both levels of the outcome shown in the output.
    - The odds of selecting the B product rather than the A product are 0.09 times lower for males than females, all else held constant (or, the odds are 91% lower)
    - The odds of selecting the C product rather than the A product are 1.1 times higher for males than females, all else held constant (or, the odds are 10% higher)

- Per year increase in age, the odds of selecting the B product compared to the A product increase 1.28 times (or, increase by 28%).
- Per year increase in age, the odds of selecting the C product compared to the A product increase 1.3 times (or, increase by 30%).
- Per additional household member, the odds of selecting the B product compared to the A product are reduced by 0.38 times (or, decrease by 62%).
- Per additional household member, the odds of selecting the C product compared to the A product increase 1.23 times (or, increase by 23%).

4. Use the code below to generate the p-values for the coefficient estimates. Which factors are statistically significantly associated with the choice of healthcare plan? Are there any cases where a factor is significant for one level compared to the baseline but not for the other level compared to the baseline? (Note: Remember that we never report a p-value of 0! Use something like "p<0.0001")

```
z <- summary(healthinsurance_mod1)$coefficients/summary(healthinsurance_mod1)$standard.errors

(p <- (1-pnorm(abs(z)))*2)
```

```
  (Intercept) age gender_facMale gender_facNon-binary    household
B           0   0      0.0000000            0.8369465 0.000000e+00
C           0   0      0.6207192            0.5323278 3.796088e-05
  position_level    absent
B   3.199529e-06 0.3684170
C   9.419906e-03 0.7926958
```

```
data.frame(t(p))
```

```
                                B            C
(Intercept)          0.000000e+00 0.000000e+00
age                  0.000000e+00 0.000000e+00
gender_facMale       0.000000e+00 6.207192e-01
gender_facNon-binary 8.369465e-01 5.323278e-01
household            0.000000e+00 3.796088e-05
position_level       3.199529e-06 9.419906e-03
absent               3.684170e-01 7.926958e-01
```

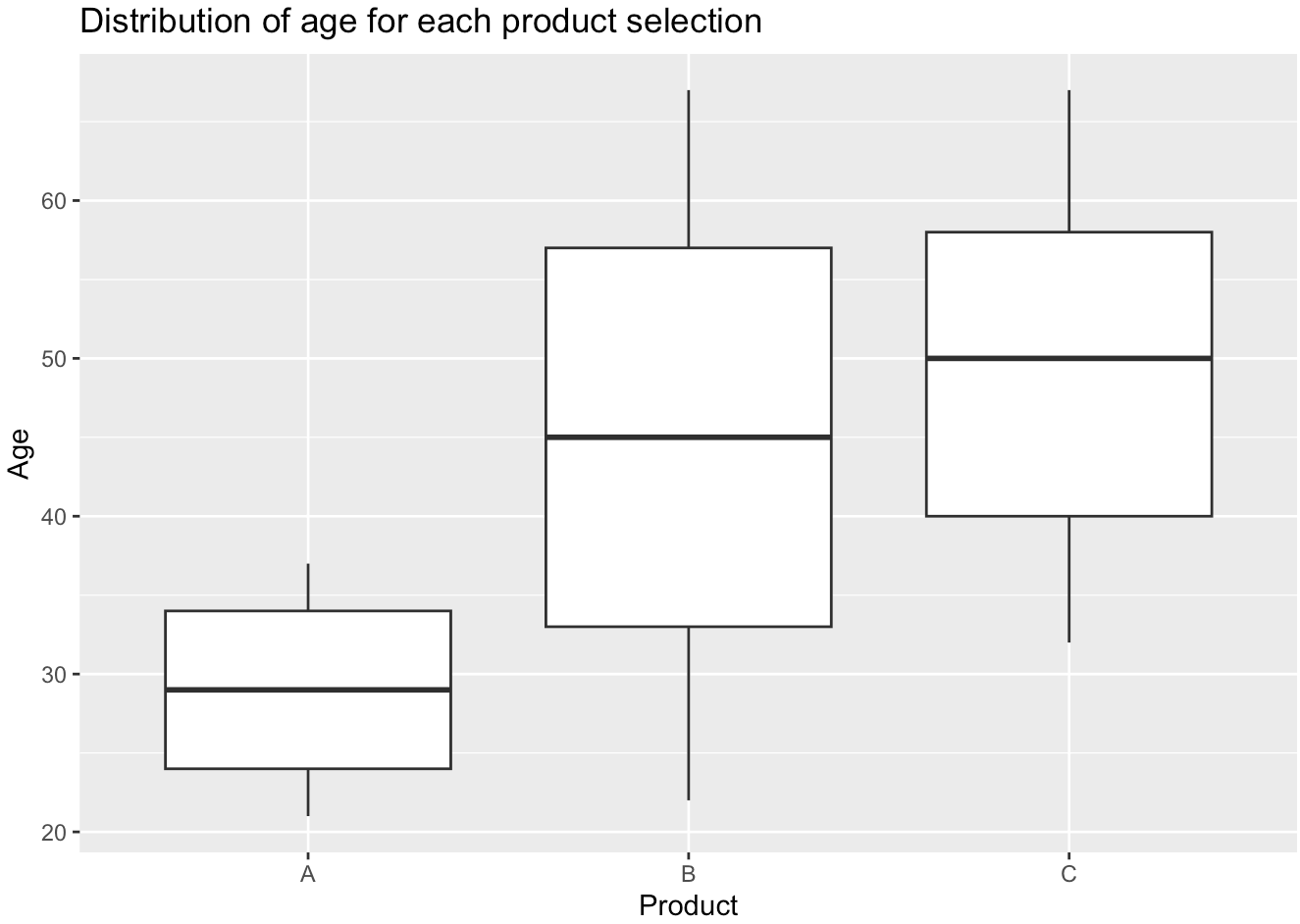
Age, household, and position level are significantly associated with product choice for both B compared to A and C compared to A. The difference in odds for males vs females is statistically significant for B compared to A but not

for C compared to A.

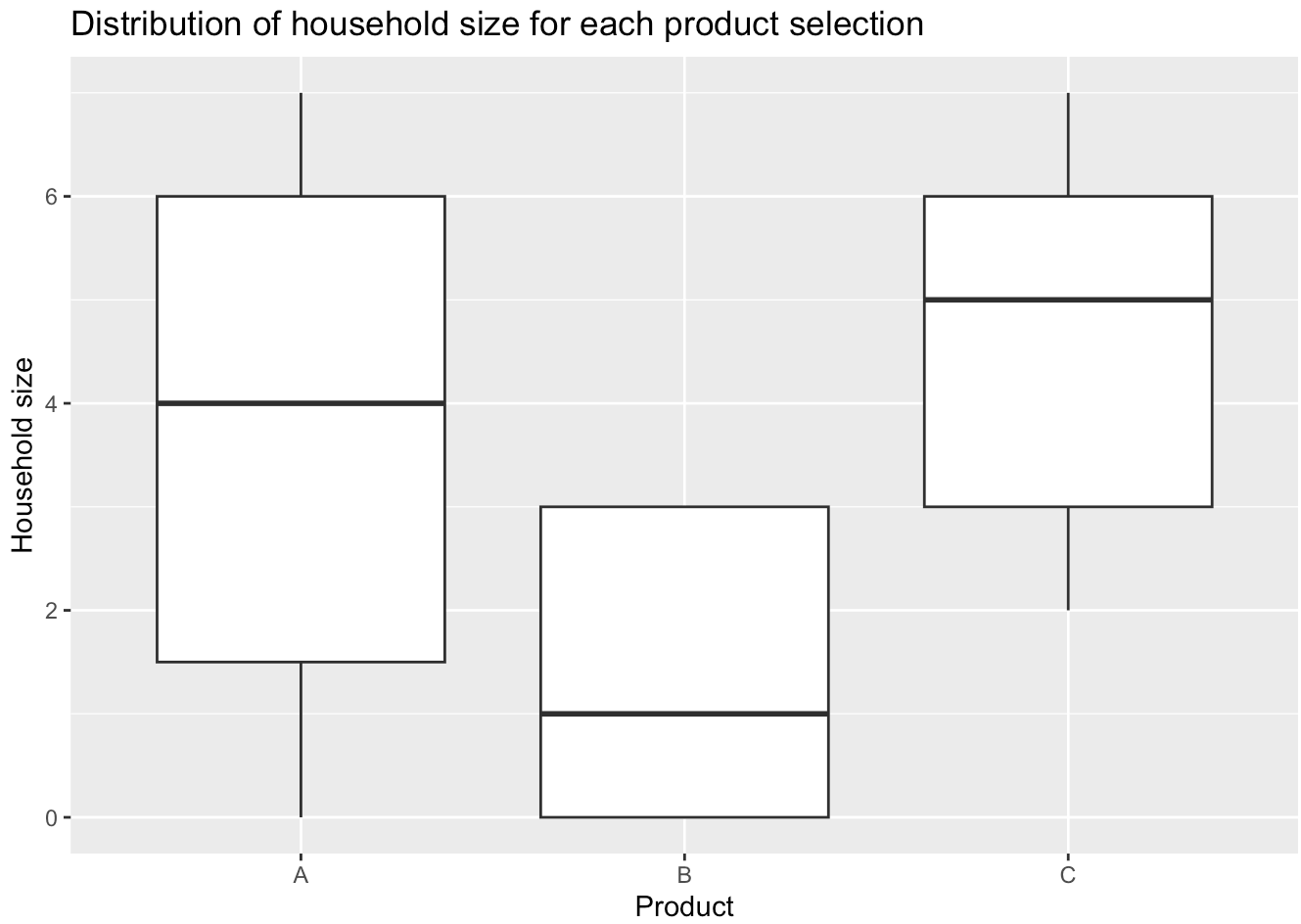
5. Generate a plot or two to illustrate some of the compelling results.

   Here are some options to illustrate the difference in selected product based on age, household size, and gender:

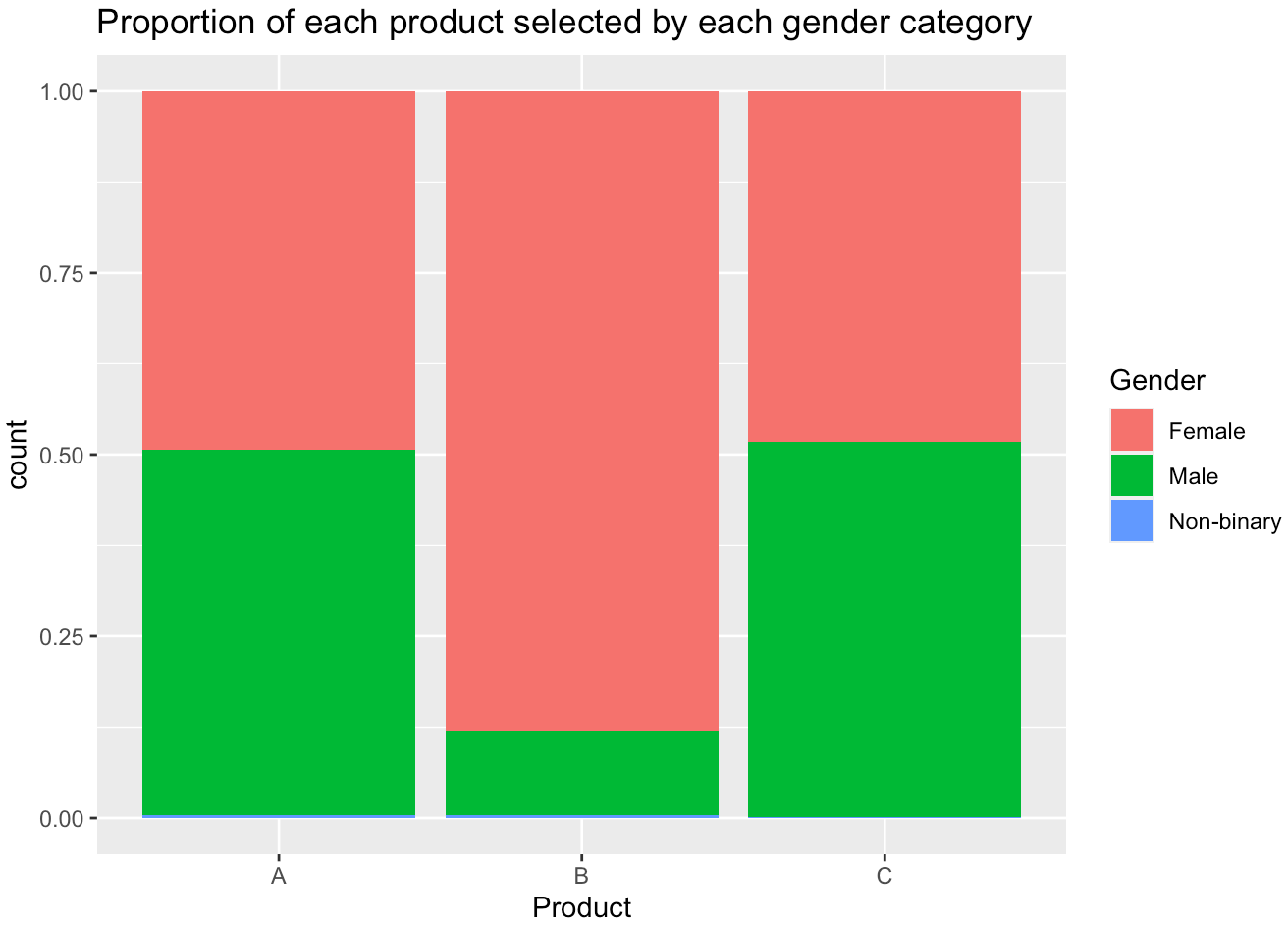
```
ggplot(health_insurance, aes(x=product_fac, y=age))+
  geom_boxplot()+
  labs(x="Product", y="Age", title="Distribution of age for each product selection")
```




```
ggplot(health_insurance, aes(x=product_fac, y=household))+
  geom_boxplot()+
  labs(x="Product", y="Household size", title="Distribution of household size for each product selectio
```




```
ggplot(health_insurance, aes(x=product_fac, fill=gender_fac))+
  geom_bar(position="fill")+
  labs(x="Product", fill="Gender", title="Proportion of each product selected by each gender category")
```



Reference

Exercise adapted from https://peopleanalytics-regression-book.org/multinomial-logistic-regression-for-nominal-category-outcomes.html