Week 8: Logistic regression

Monica Alexander

06/03/2022

Overview

We are going to look at the association of income (low income or not), age and education in the GSS.

Data

Load in the GSS

```
library(tidyverse)
library(here)
gss <- read_csv(here("data/gss.csv"))
```

Create binary outcome variable

The income variable in the GSS is categorical. You can look at different categories by using the `unique` function:

```
unique(gss$income_respondent)
```

```
## [1] "$25,000 to $49,999"    "Less than $25,000"     "$50,000 to $74,999"
## [4] "$125,000 and more"     "$75,000 to $99,999"    "$100,000 to $ 124,999"
```

For the purposes of our logistic regression, let's define a binary variable called "low income" that is equal to 1 if a respondent is in the less than $25,000 category and 0 otherwise.

```
gss <- gss %>%
  mutate(low_income = ifelse(income_respondent=="Less than $25,000", 1, 0))
```

NOTE: this is not the only possible way of studying income as a binary outcome. For example, you may be more interested in whether or not a respondent has high income, and define a binary variable called `high_income` if the respondent has more than $125,000. Additionally, you can define a binary outcome based on more than one category. For example, if we were interested in whether respondents earned more or less than $50,000, we could define the following:

```
gss <- gss %>%
  mutate(income_less_than_50k = ifelse(income_respondent=="Less than $25,000"|
                                         income_respondent=="$25,000 to $49,999", 1, 0))
```

Note that the vertical bar "|" means "or" so you read the above code as "if the income of the respondent is less than $25,000 OR between $25,000 and $49,000 then the new variable `income_less_than_50k` is equal to 1, otherwise it is equal to 0."

Logistic regression

Now run a logistic regression with dependent variable `low_income` and independent variable age.

```
mod <- glm(low_income ~ age, family = "binomial", data = gss)
summary(mod)
```

```
##
## Call:
## glm(formula = low_income ~ age, family = "binomial", data = gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9684  -0.9049  -0.8653   1.4361   1.5630
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.421924   0.045567  -9.260  < 2e-16 ***
## age         -0.005630   0.000835  -6.742 1.56e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26093  on 20601  degrees of freedom
## Residual deviance: 26047  on 20600  degrees of freedom
## AIC: 26051
##
## Number of Fisher Scoring iterations: 4
```

Remember that the coefficients are on the 'log-odds' scale. To convert to the odds scale, you can exponentiate:

```
exp(coef(mod))
```

```
## (Intercept)         age
##   0.6557838   0.9943861
```

Question

Interpret β_1 and $\exp \beta_1$

Regression with age groups

using age as a quantitative variable as above assumes that the association with income and age is always constant: that is, the probability of low income always decreases with increased age. But do we believe this? It might be the case that the likelihood of low income changes over age profiles. To investigate this, we can define a age group categorical variable and run a regression with this variable.

First define a 10 year age group variable:

```
age_groups <- seq(10, 80, by = 10)

gss$age_group <- as.character(cut(gss$age,
                    breaks= c(age_groups, Inf),
                    labels = age_groups,
                    right = FALSE))
```

Now run a regression:

```
mod_2 <- glm(low_income ~ age_group, family = "binomial", data = gss)
summary(mod_2)
```

```
##
## Call:
## glm(formula = low_income ~ age_group, family = "binomial", data = gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8477  -0.8950  -0.6991   1.2373   1.8240
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.0372     0.2912   13.87   <2e-16 ***
## age_group20  -4.1770     0.2945  -14.18   <2e-16 ***
## age_group30  -5.3216     0.2942  -18.09   <2e-16 ***
## age_group40  -5.4905     0.2950  -18.61   <2e-16 ***
## age_group50  -5.0774     0.2936  -17.30   <2e-16 ***
## age_group60  -4.7453     0.2930  -16.19   <2e-16 ***
## age_group70  -4.4094     0.2938  -15.01   <2e-16 ***
## age_group80  -4.4701     0.2972  -15.04   <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: 26093  on 20601  degrees of freedom
## Residual deviance: 23871  on 20594  degrees of freedom
## AIC: 23887
##
## Number of Fisher Scoring iterations: 6
```

Questions

- What is the reference category?
- Interpret the coefficients on age group = 20 and age group = 50. What does this suggest?

Changing reference category

Now rerun the regression based on the re-leveled age group:

```
gss <- gss %>%
  mutate(age_group = fct_relevel(age_group, "30", after = 0))

mod_3 <- glm(low_income ~ age_group, family = "binomial", data = gss)
summary(mod_3)
```

```
##
## Call:
## glm(formula = low_income ~ age_group, family = "binomial", data = gss)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8477  -0.8950  -0.6991   1.2373   1.8240
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.28439    0.04250 -30.223  < 2e-16 ***
## age_group10  5.32157    0.29425  18.085  < 2e-16 ***
## age_group20  1.14453    0.06150  18.610  < 2e-16 ***
## age_group40 -0.16898    0.06347  -2.662  0.00776 **
## age_group50  0.24415    0.05658   4.315  1.6e-05 ***
## age_group60  0.57626    0.05390  10.691  < 2e-16 ***
## age_group70  0.91222    0.05794  15.744  < 2e-16 ***
## age_group80  0.85146    0.07305  11.656  < 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: 26093  on 20601  degrees of freedom
## Residual deviance: 23871  on 20594  degrees of freedom
## AIC: 23887
##
## Number of Fisher Scoring iterations: 6
```

Questions

- Interpret the coefficients on age group = 20 and age group = 50. Why do these differ?

Further exercises

- Rerun the regression above (mod_3) with `educ_cat` as an additional explanatory variable. What is the reference category? Interpret some of the results.
- Change the reference category for education to be high school and rerun the above regression.