Test 04: Logit Regressions

Richard Dien Giau Bui

Load required packages

```
library(tidyverse)
# if you're using macOS, you can run: library(dplyr)
library(skimr)
library(broom)
```

When the dependent variable is a dummy

- Dummy: a variable receives only two possible values: 1 or 0
- Examples:
  - Pass (1) and Fail (0)
  - Got COVID (1) and not (0)
  - Got accepted to a top university (1) and a lower one (0)
  - A firm filed bankruptcy (1) and others (0)

Let's use Hsb data again as an example

```
Hsb = read_csv("data/raw/hsb.csv")
Hsb = Hsb %>%
  mutate(
    race = as.factor(race),
    schtyp = as.factor(schtyp),
    prog = as.factor(prog)
  )
```

Run an OLS regression with a dummy dependent variable

- Take the `female`: =1 for female student and =0 for male student
- Imagine we want to train data so that if we know the student's reading, writing, math, and science scores, the model will guess if the student is female or male student
- How to run this model by OLS regression?

OLS

```r
m1 = lm(female ~ read + write + math + science, data=Hsb[1:150,])
summary(m1)
```

```
Call:
lm(formula = female ~ read + write + math + science, data = Hsb[1:150,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.8945 -0.3976 -0.1468  0.4616  0.8847

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.304953   0.234283   1.302  0.19510
read        -0.005086   0.005246  -0.970  0.33385
write        0.025934   0.005586   4.643 7.62e-06 ***
math        -0.003932   0.005684  -0.692  0.49025
science     -0.014042   0.005059  -2.776  0.00626 **
```

Discussion

- We have a model with meaningful coefficients,
- Everything is fine
- Like a movie/novel, every bad thing happens with a **BUT**

> … **BUT**, let's check the fitted value

Question

- Remember which function to get the fitted and residual from a regression?

augment

```
augment(m1, newdata = Hsb[151:200,])
```

```
# A tibble: 50 × 13
      id female race    ses schtyp prog   read write  math science socst
.fitted
   <dbl>  <dbl> <fct> <dbl> <fct>  <fct> <dbl> <dbl> <dbl>   <dbl> <dbl>
<dbl>
 1    43      1 3         1 1      2        47    37    43      42    46
0.229
 2    96      1 4         3 1      2        65    54    61      58    56
0.268
 3   138      1 4         2 1      3        43    57    40      50    51
0.660
 4    10      1 1         2 1      1        47    54    49      53    61
0.482
 5    71      1 4         2 1      1        57    62    56      58    66
0.536
 6   120      1 4         2 1      2        68    50    61      55    71
```

Discussion

- Do you notice anything unconventional in the `.fitted` column?

Logit regression

- When facing the dummy dependents, it is better to use logit regression
- In R, we use the function `glm`

```
Call:
glm(formula = female ~ read + write + math + science, family = "binomial",
    data = Hsb)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.51473    1.01924  -0.505  0.61355
read        -0.03001    0.02296  -1.307  0.19116
write        0.15295    0.02790   5.482  4.2e-08 ***
math        -0.02962    0.02608  -1.136  0.25607
science     -0.08160    0.02507  -3.255  0.00114 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

augment again

```
augment(m2, newdata = Hsb[151:200,])
```

```
# A tibble: 50 × 12
      id female race    ses schtyp prog   read write  math science socst
.fitted
   <dbl>  <dbl> <fct> <dbl> <fct>  <fct> <dbl> <dbl> <dbl>   <dbl> <dbl>
<dbl>
 1    43      1 3         1 1      2        47    37    43      42    46
-0.967
 2    96      1 4         3 1      2        65    54    61      58    56
-0.746
 3   138      1 4         2 1      3        43    57    40      50    51
1.65
 4    10      1 1         2 1      1        47    54    49      53    61
0.558
 5    71      1 4         2 1      1        57    62    56      58    66
0.866
```

Note

- The logit formula is not standard as the OLS, so we don't care much about the range of the fitted value anymore
- Simply just look at the sign (positive or negative) of the coefficients, rather than the size of coefficients
- The details of math behind the logit regression is skipped for simplicity in this class
  - If you want to learn more, maybe you can read yourself. My recommendation is Wooldridge textbook on Introduction Econometrics

Logistic regression as a classification tool in machine learning

Next

- You’re armed with many statistical tools now
- So the analysis system will be:
  - Clean data
  - Statistical tests
  - … then report the results to me
- Next lectures will focus on how to make good reporting results to audience