Week3 Lab 134

1. Obtain dataset *Discrim*

(a) Dataset description and some EDA

Discrim is a simulated dataset containing $n = 28$ job interview outcomes of a company on $p = 4$ features.

- `HIRING`: response variable with two levels, "1" stands for YES and "0" for NO
- `EDUCATION`: years of college education, three values are available
- `EXPERIENCE`: years of working experience
- `GENDER`: "1" for MALE and "0" for FEMALE

```r
library(tidyverse)
# Read the txt file from your current working directory
Dis = read.table("Discrim.txt", header=T)
# Convert Dis into a data frame
Dis = as_tibble(Dis)
str(Dis)
```

```
## tibble [28 x 4] (S3: tbl_df/tbl/data.frame)
##  $ HIRING    : int [1:28] 0 0 1 1 0 1 0 0 0 1 ...
##  $ EDUCATION : int [1:28] 6 4 6 6 4 8 4 4 6 8 ...
##  $ EXPERIENCE: int [1:28] 2 0 6 3 1 3 2 4 1 10 ...
##  $ GENDER    : int [1:28] 0 1 1 1 0 0 1 0 0 0 ...
```

Convert categorical variables to factor since glm() treats them as numeric otherwise.

```
```r
# install.packages("dplyr")
library(dplyr)
Dis = Dis %>%
  mutate(HIRING=as.factor(ifelse(HIRING==0,"No", "Yes"))) %>%
  mutate(GENDER=as.factor(ifelse(GENDER==0,"F", "M")))
str(Dis)
```

```
## tibble [28 x 4] (S3: tbl_df/tbl/data.frame)
##  $ HIRING    : Factor w/ 2 levels "No","Yes": 1 1 2 2 1 2 1 1 1 2 ...
##  $ EDUCATION : int [1:28] 6 4 6 6 4 8 4 4 6 8 ...
##  $ EXPERIENCE: int [1:28] 2 0 6 3 1 3 2 4 1 10 ...
##  $ GENDER    : Factor w/ 2 levels "F","M": 1 2 2 2 1 1 2 1 1 1 ...
```

Let's check some explanatory analysis on the dataset.


```r
table(Dis$GENDER,Dis$HIRING)
```
```

```

```

```
    No Yes
  F 12   3
  M  7   6
```

```
- Among 15 FEMALE applying, 3 have been hired.

- Among 13 MALE applying, 6 have been hired.
```

(b) Interesting questions

Based on the dataset, we may pose some intriguing questions like

- Why is a logistic regression model better than a linear one?

- What is the probability of being hired given some features of candidates (`EDUCATION`, `EXPERIENCE` and `GENDER` of a candidate)?

- Does each predictor actually have impact on the estimated probabilities in the logistic model?

2. Logistic Regression

(a) Review the theoretical background

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta' X \Longleftrightarrow p(Y = j \mid X) = \frac{e^{\beta' X}}{1 + e^{\beta' X}}$$

where $\beta' X = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

(b) Build and summarise a logistic regression model

- **`glm()`** is used to fit generalized linear models. The usage of `glm()` is pretty much like that of `lm()` with one more necessary argument `family`. Specifying **`family=binomial`** produces a logistic regression model. By default, `family=binomial` uses logit as its link function. More options such as probit, log-log link are also available. As described previously, `HIRING` is our response and `EDUCATION`, `EXPERIENCE` and `GENDER` are predictors.

- **`summary()`** is a generic function that is used to produce result summaries of various model fitting functions. We can call the `summary()` of our glm object after fitting it and expect several things to be reported:

  - *Call*: this is R reminding us what the model we ran was, what options we specified, etc

  - *Deviance residuals*: measures of model fit. This part of output shows the distribution of the deviance residuals for individual cases used in the model

  - *Coefficients*: shows the coefficients, their standard errors, the Z-statistic (sometimes called a Wald Z-statistic), and the associated p-values

  - *Fit indices*: goodness-of-fit measures including the null and deviance residuals, and the AIC.

```r
# Specify 'family=binomial' is important!
glm.fit = glm(HIRING ~ EDUCATION + EXPERIENCE + GENDER,
             data=Dis, family=binomial)
```

```r
# Summarize the logistic regression model
summary(glm.fit)
```

```
##
## Call:
## glm(formula = HIRING ~ EDUCATION + EXPERIENCE + GENDER, family = binomial,
##     data = Dis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4380  -0.4573  -0.1009   0.1294   2.1804
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -14.2483     6.0805  -2.343   0.0191 *
## EDUCATION     1.1549     0.6023   1.917   0.0552 .
## EXPERIENCE    0.9098     0.4293   2.119   0.0341 *
## GENDERM       5.6037     2.6028   2.153   0.0313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35.165  on 27  degrees of freedom
## Residual deviance: 14.735  on 24  degrees of freedom
## AIC: 22.735
##
## Number of Fisher Scoring iterations: 7
```

(c) Interpret coefficients

In above results, Both `EXPERIENCE` and `GENDERM` are statistically significant at level 0.05.

Let's take a look at the interpretation of the model coefficients. So, our model is

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 * Education + \beta_2 * Experience + \beta_3 * Gender$$

where p = probability of getting hired.

The logistic regression coefficients, if logit link function is used, give the change in the log odds of the outcome for a one unit increase in a predictor variable, while others being held constant.

- The variable `EXPERIENCE` has a coefficient 0.9098. For every one unit change in `EXPERIENCE`, the log odds of getting hired (versus not-hired) increases by 0.9098, holding other variables fixed. Mathematically, after 1 unit increase of experience,

$$log(\frac{p_{new}}{1-p_{new}}) = 0.9098 + log(\frac{p_{old}}{1-p_{old}})$$

- The variable `EDUCATION` has a coefficient 1.1549. For a one unit increase in `EDUCATION`, the log odds of being hired increases by 1.1549, holding other variables fixed
- The indicator variable for `GENDERM` has a slightly different interpretation. The variable `GENDERM` has a coefficient 5.6037, meaning that the indicator function of `MALE` has a regression coefficient 5.6037. That being said, the gender `MALE` versus `FEMALE`, changes the log odds of getting hired by 5.6037.

Poisson regression

- For count/rate data

Examples:

- Number of cargo ships damaged by waves (classic example given by McCullagh & Nelder, 1989).
- Daily homicide counts in California (Grogger, 1990)
- Number of arrests resulting from 911 calls.

(a) Model

- Response: Poisson distribution and model the expected value of Y, denoted by $E(Y) = \mu$.
- Predictors : For now, just 1 explanatory variable x as example.
- Link: We could use
- Identity link, which gives us $\mu = \alpha + \beta x$

Problem: a linear model can yield $\mu < 0$, while the possible values for $\mu \geq 0$

- Log link (much more common) $\log(\mu)$, which is the "natural parameter" of Poisson distribution, and the log link is the "canonical link" for GLMs with Poisson distribution.

The Poisson regression model for counts (with a $\log \text{link}$) is

$$\log(\mu) = \alpha + \beta x$$

This is often referred to as "Poisson loglinear model".

For this single variate poisson, let's see how does 1 unit change in predictor affects response (count).

$$\log(\mu) = \beta_0 + \beta_1 x$$

Consider distinct $x\,(x_1 \& x_2)$ such that the difference between them equals 1 . For example, $x_1 = 10$ and $x_2 = 11$:

$$x_2 = x_1 + 1$$

The expected value of μ when $x = 10$ is

$$\mu_1 = e^{\alpha} e^{\beta x_1} = e^{\alpha} e^{\beta(10)}$$

The expected value of μ when $x = x_2 = 11$ is

$$\begin{aligned}
\mu_2 &= e^{\alpha} e^{\beta x_2} \\
&= e^{\alpha} e^{\beta(x_1+1)} \\
&= e^{\alpha} e^{\beta x_1} e^{\beta} \\
&= e^{\alpha} e^{\beta(10)} e^{\beta} = \mu_1 e^{\beta}
\end{aligned}$$

A change in x has a multiplicative effect on the mean of Y.

Case 1: If $\beta = 0$, then $e^0 = 1$ and

- $\mu_1 = e^{\alpha}$.
- $\mu_2 = e^{\alpha}$.
- $\mu = E(Y)$ is not related to x.

Case 2: If $\beta > 0$, then $e^{\beta} > 1$ and

- $\mu_1 = e^{\alpha} e^{\beta x_1}$
- $\mu_2 = e^{\alpha} e^{\beta x_2} = e^{\alpha} e^{\beta x_1} e^{\beta} = \mu_1 e^{\beta}$
- μ_2 is e^{β} times larger than μ_1.

Case 3: If $\beta < 0$, then $0 \leq e^{\beta} < 1$

- $\mu_1 = e^{\alpha} e^{\beta x_1}$.
- $\mu_2 = e^{\alpha} e^{\beta x_2} = e^{\alpha} e^{\beta x_1} e^{\beta} = \mu_1 e^{\beta}$.
- μ_2 is e^{β} times smaller than μ_1.

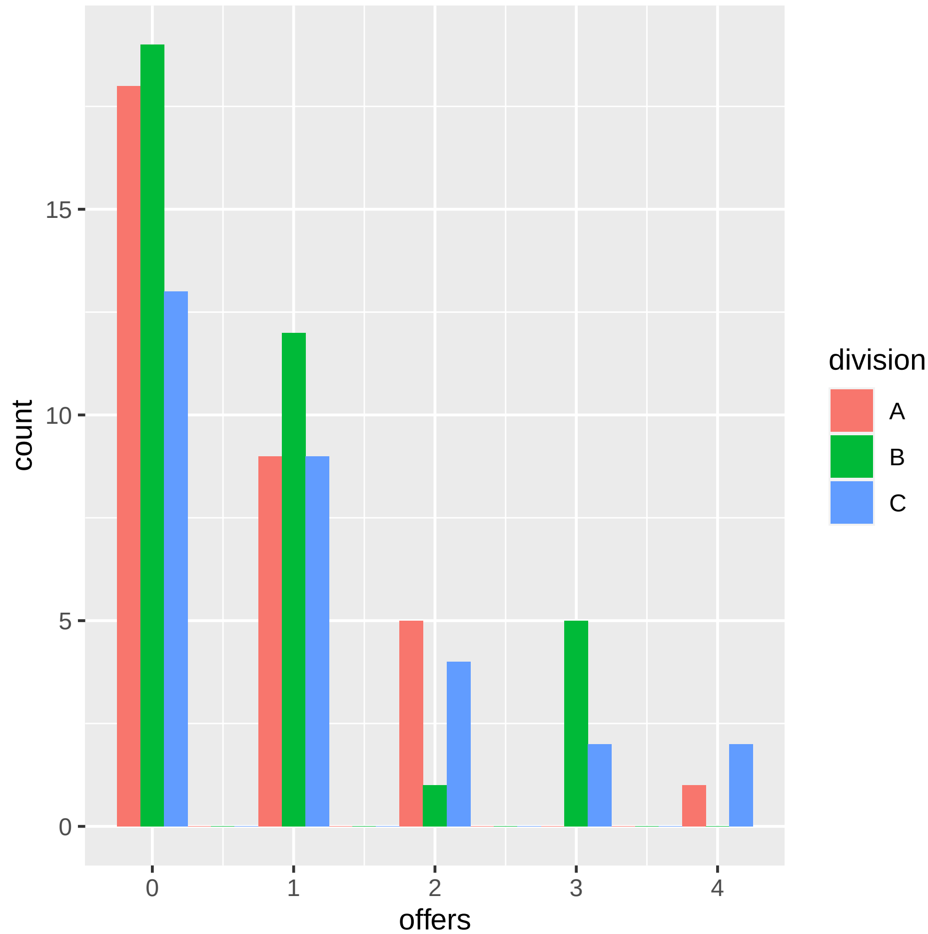
(b) Example of model

Suppose we want to know how many scholarship offers a high school baseball player in a given county receives based on their school division ("A", "B", or "C") and their college entrance exam score (measured from 0 to 100).

```
#make this example reproducible
set.seed(1)

#create dataset
data <- data.frame(offers = c(rep(0, 50), rep(1, 30), rep(2, 10), rep(3, 7), rep(4, 3)),
                   division = sample(c("A", "B", "C"), 100, replace = TRUE),
                   exam = c(runif(50, 60, 80), runif(30, 65, 95), runif(20, 75, 95)))

#create histogram
library(ggplot2)
ggplot(data, aes(offers, fill = division)) +
  geom_histogram(binwidth=.5, position="dodge")
```



Above is a visualization of number of offers received by players based on division. We see most players receive either 0 or 1 offer.

Let's fit the model and interpret some coefficents.

```
#fit the model
model <- glm(offers ~ division + exam, family = "poisson", data = data)

#view model output
summary(model)
```

```
##
## Call:
## glm(formula = offers ~ division + exam, family = "poisson", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3376  -0.8612  -0.6167   0.2442   2.6496
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.21183    1.04877  -6.876 6.14e-12 ***
## divisionB    0.07156    0.27935   0.256    0.798
## divisionC    0.26906    0.27585   0.975    0.329
```

```
## exam         0.08614    0.01236   6.969 3.20e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 138.069  on 99  degrees of freedom
## Residual deviance:  82.741  on 96  degrees of freedom
## AIC: 207.62
##
## Number of Fisher Scoring iterations: 5
```

(c) Interpreting coefficients

The coefficient for exam is 0.08614. i.e; The expected log count for number of offers for a one-unit increase in exam is 0.08614. An easier way to interpret this is to take the exponent as in (b), that is $e^{0.08614} = 1.09$. So, there is a 9% increase in the number of offers received for each additional point scored on the entrance exam.

Let's look at coefficient for division B, 0.07156. Take exponent, $e^{0.07156} = 1.07$ which means players in divisionB receive 7% more offers than players in division A. Note the difference is not significant (p-value >0.05).

Similary, for division C, we have $e^{0.26906} = 1.309$ which means players in division C receive more offer than players in division A by 30%. Again, not significant (p-value >0.05).

3 GLMM

Example : A large HMO wants to know what patient and physician factors are most related to whether a patient's lung cancer goes into remission after treatment as part of a larger study of treatment outcomes and quality of life in patients with lunger cancer. A variety of outcomes were collected on patients, who are nested within doctors, who are in turn nested within hospitals.

Below we use the glmer command to estimate a mixed effects logistic regression model with Il6, CRP, and LengthofStay as patient level continuous predictors, CancerStage as a patient level categorical predictor (I, II, III, or IV), Experience as a doctor level continuous predictor, and a random intercept by DID, doctor ID.

```r
hdp <- read.csv("https://stats.idre.ucla.edu/stat/data/hdp.csv")
hdp <- within(hdp, {
  Married <- factor(Married, levels = 0:1, labels = c("no", "yes"))
  DID <- factor(DID)
  HID <- factor(HID)
  CancerStage <- factor(CancerStage)
})
#install.packages("lme4")
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 4.2.1

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```r
# estimate the model and store results in m
m <- glmer(remission ~ IL6 + CRP + CancerStage + LengthofStay + Experience +
             (1 | DID), data = hdp, family = binomial)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.250042 (tol = 0.002, component 1)
```

```r
# print the mod results without correlations among fixed effects
summary(m)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: remission ~ IL6 + CRP + CancerStage + LengthofStay + Experience +
##     (1 | DID)
##    Data: hdp
##
##      AIC      BIC   logLik deviance df.resid
##   7410.1   7473.6  -3696.1   7392.1     8516
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7571 -0.4426 -0.2017  0.3986  7.0988
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  DID    (Intercept) 3.894    1.973
## Number of obs: 8525, groups:  DID, 407
##
## Fixed effects:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.029592   0.514568  -3.944 8.00e-05 ***
## IL6            -0.055678   0.011243  -4.952 7.34e-07 ***
## CRP            -0.020531   0.009981  -2.057 0.039678 *
## CancerStageII  -0.413356   0.073882  -5.595 2.21e-08 ***
## CancerStageIII -1.000274   0.095998 -10.420  < 2e-16 ***
## CancerStageIV  -2.341861   0.155863 -15.025  < 2e-16 ***
## LengthofStay   -0.119944   0.032858  -3.650 0.000262 ***
## Experience      0.117975   0.026615   4.433 9.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) IL6    CRP    CncSII CnSIII CncSIV LngthS
## IL6         -0.084
## CRP         -0.088  0.002
## CancerStgII  0.014  0.006  0.005
## CancrStgIII  0.057  0.008  0.015  0.493
## CancerStgIV  0.065  0.030  0.013  0.332  0.317
## LengthofSty -0.300  0.012 -0.020 -0.265 -0.337 -0.288
## Experience  -0.915 -0.005 -0.002 -0.004 -0.008 -0.013 -0.010
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.250042 (tol = 0.002, component 1)
```

References:

https://stats.oarc.ucla.edu/r/dae/mixed-effects-logistic-regression/

https://www.statology.org/poisson-regression/