# Lab 3: Logistic Regression (Part 1)

**<u>To submit your work, insert screenshots of your code and outputs (both numeric outputs and graphs) under respective problem prompts. Many steps also require a written answer, and you should insert your written or typed answer below the prompt.</u>**

*Suppose you are investigating allegations of gender discrimination in the hiring practices of a particular firm. An equal-rights group claims that females are less likely to be hired than males with the same background, experience, and other qualifications. You collected data on 28 former applicants. The variables in the dataset include:*

- *HIRE (1 = hired, 0 = not hired)*
- *Years of higher education (EDUC)*
- *Years of work experience (EXP)*
- *GENDER (1 = male, 0 = female).*

1. Download the data file “DISCRIM.csv” from Canvas.

   DISCRIM.csv 11/18/2023 9:00 PM

2. Start R or R Studio. Load the “car” package.

```
rm(list=ls())
setwd("C:/Users/jyang/OneDrive - Arizona State University/10 Classes_OneDrive/2023_STP530_Regression/R")
library(car)
# install.packages("caret")
library(ggplot2)
library(lattice)
library(caret)
# install.packages("OptimalCutpoints")
library(OptimalCutpoints)
```

```
> library(car)
Loading required package: carData
> # install.packages("caret")
> library(ggplot2)
> library(lattice)
> library(caret)
> # install.packages("OptimalCutpoints")
> library(OptimalCutpoints)
```

3. Import the data into R. Name the imported data **hire.data**. View the data and make sure the data have been imported correctly.

```
# Load data
hire.data <- read.csv("DISCRIM.csv")

# Check data
head(hire.data )
str(hire.data )
```

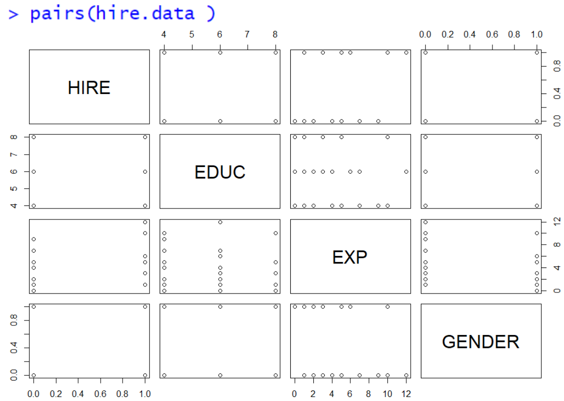
```
> # Load data
> hire.data <- read.csv("DISCRIM.csv")
>
> # Check data
> head(hire.data )
  HIRE EDUC EXP GENDER
1    0    6   2      0
2    0    4   0      1
3    1    6   6      1
4    1    6   3      1
5    0    4   1      0
6    1    8   3      0
> str(hire.data )
'data.frame':   28 obs. of  4 variables:
 $ HIRE  : int  0 0 1 1 0 1 0 0 0 1 ...
 $ EDUC  : int  6 4 6 6 4 8 4 4 6 8 ...
 $ EXP   : int  2 0 6 3 1 3 2 4 1 10 ...
 $ GENDER: int  0 1 1 1 0 0 1 0 0 0 ...
```

4. Inspect the variables individually and pairwise. <u>Describe your impression.</u>

```
summary(hire.data)

table(hire.data$HIRE)

table(hire.data$GENDER)

pairs(hire.data)
```

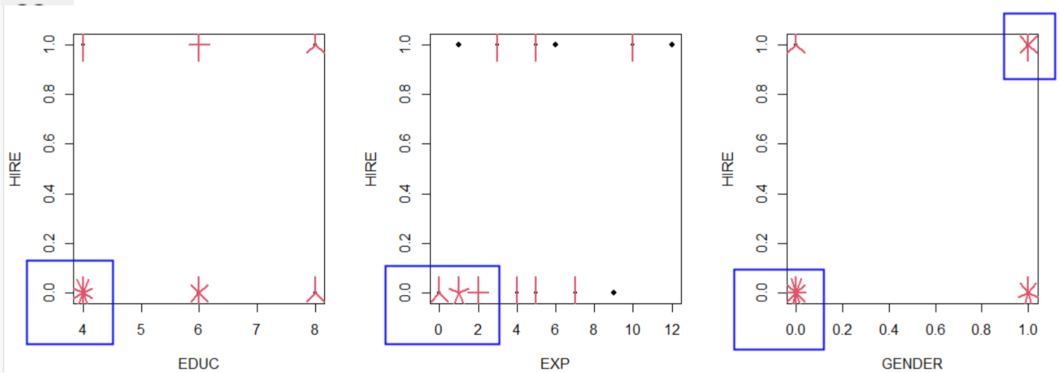
```
summary(hire.data )
table(hire.data $HIRE)
table(hire.data $GENDER)
pairs(hire.data )
```

```
> summary(hire.data )
      HIRE             EDUC            EXP             GENDER      
 Min.   :0.0000   Min.   :4.000   Min.   : 0.000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:4.000   1st Qu.: 1.000   1st Qu.:0.0000  
 Median :0.0000   Median :6.000   Median : 3.000   Median :0.0000  
 Mean   :0.3214   Mean   :5.571   Mean   : 3.893   Mean   :0.4643  
 3rd Qu.:1.0000   3rd Qu.:6.000   3rd Qu.: 5.250   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :8.000   Max.   :12.000   Max.   :1.0000  
> table(hire.data $HIRE)

 0  1 
19  9 
> table(hire.data $GENDER)

 0  1 
15 13 
> pairs(hire.data )
```



HIRE and GENDER are binary,
EDUC and EXP are ordinal. No immediately obvious linear relationships is shown.

```
par(mfrow=c(1,3))
with(hire.data , sunflowerplot(EDUC, HIRE))
with(hire.data , sunflowerplot(EXP, HIRE))
with(hire.data , sunflowerplot(GENDER, HIRE))
```




Sunflowerplot show a potential relationship between EDUC&HIRE, EXP&HIRE, and GENDER&HIRE, particularly, concentrated spots at not-hired with low-education, low-experience, and female.

5. Fit a logistic regression model.

```
m <- glm(HIRE ~ EDUC + EXP + GENDER, data=hire.data,
               family=binomial)
summary(m)
```

```
# Fit logistic regression model
m <- glm(HIRE ~ EDUC + EXP + GENDER, data=hire.data , family=binomial)
summary(m)
```

```
> # Fit logistic regression model
> m <- glm(HIRE ~ EDUC + EXP + GENDER, data=hire.data , family=binomial)
> summary(m)

Call:
glm(formula = HIRE ~ EDUC + EXP + GENDER, family = binomial,
    data = hire.data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.2483     6.0805  -2.343   0.0191 *
EDUC          1.1549     0.6023   1.917   0.0552 .
EXP           0.9098     0.4293   2.119   0.0341 *
GENDER        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
```

6. Write out the fitted model (original form) with the estimated coefficient values and meaningful variable names.

Pi = probability of hired

Odds = $\frac{P(sucess)}{P(failure)}$ = $\frac{pi}{1-pi}$ , E{HIRE} = pi

Logit(HIRE) = log-odd{HIRE}= $ln\,\frac{pi}{1-pi}$

= -14.2483 + 1.1549(EDUC) + 0.9098(EXP) + 5.6037(GENDER.male)

7. **Interpret model coefficients.** Interpret each of three slope coefficients in at least two ways. Refer to the lecture slides for different ways to interpret slope coefficients.

   a. Higher education (b1 = 1.1549)

   **Interpretation 1)** Holding years of experience and gender constant, the predicted log-odds that the candidate gets hired increases by 1.1549 with every 1 more year of higher education.

   **Interpretation 2)** e^1.1549: Holding years of experience and gender constant, the predicted odds that the candidate gets hired increases by a factor of 3.1737 (=e^1.1549) with every 1 more year of higher education.

   **Interpretation 3)** Holding years of experience and gender constant, the predicted odds that the candidate gets hired increases by 217.37% (e^1.1549 – 1, %) with every 1 more year of higher education.

   b. Work experience (b2 = 0.9098)

   **Interpretation 1)** Holding years of higher education and gender constant, the predicted log-odds that the candidate gets hired increases by 0.9098 with every 1 more year of experience.

   **Interpretation 2)** Holding years of higher education and gender constant, the predicted odds that the candidate gets hired increases by a factor of 2.4838 (=e^0.9098) with every 1 more year of experience.

   **Interpretation 3)** Holding years of higher education and gender constant, the predicted odds that the candidate gets hired increases by 148.38% (e^0.9098– 1, %) with every 1 more year of higher education.

   c. Gender (b3 = 5.6037)

   **Interpretation 1)** Holding years of higher education and years of experience constant, the predicted log-odds that a male candidate gets hired is higher than the log-odds for a female candidate by 5.6037.

   **Interpretation 2)** Holding years of higher education and years of experience constant, the predicted odds that a male candidate gets hired is higher than the odds for a female candidate by a factor of 271.4288 (e^5.6037).

   **Interpretation 3)** Holding years of higher education and years of experience constant, the predicted odds that a male candidate gets hired is 271.4288 % (e^5.6037-1 %) times higher than a female candidate.

8. **Confidence interval of model coefficient.** Compute the 95% confidence interval of the slope coefficient of EDUC. The point estimate and the standard error are given in the model

summary output. For the distribution multiplier, use the z-distribution (standard normal) instead of t-distribution. Report and interpret the confidence interval.

```
# Confidence interval for model coefficients

summary(m)$coefficients
qnorm(p=.975) # z-distribution

# 95% CI of b_EDU, using z-distribution
1.1549 - 1.96 * 0.6023 # LL
summary(m)$coefficients[2, 1] - qnorm(.975) * summary(m)$coefficients[2, 2]

1.1549 + 1.96 * 0.6023 # UL
summary(m)$coefficients[2, 1] + qnorm(.975) * summary(m)$coefficients[2, 2]
```

```
> #-------------------------------------------------------
> # Confidence interval for model coefficients
>
> summary(m)$coefficients
               Estimate Std. Error   z value   Pr(>|z|)
(Intercept) -14.2482575  6.0805351 -2.343257 0.01911620
EDUC          1.1548804  0.6022944  1.917468 0.05517846
EXP           0.9098486  0.4292934  2.119410 0.03405584
GENDER        5.6036794  2.6027819  2.152958 0.03132200
> qnorm(p=.975) # z-distribution
[1] 1.959964
>
> # 95% CI of b_EDU, using z-distribution
> 1.1549 - 1.96 * 0.6023 # LL
[1] -0.025608
> summary(m)$coefficients[2, 1] - qnorm(.975) * summary(m)$coefficients[2, 2]
[1] -0.0255949
>
> 1.1549 + 1.96 * 0.6023 # UL
[1] 2.335408
> summary(m)$coefficients[2, 1] + qnorm(.975) * summary(m)$coefficients[2, 2]
[1] 2.335356
```

- With 95% confidence, the b1 (EDUC) falls between -0.025608 and 2.335356.
- With 95% confidence, we estimate that the log-odds of success (being hired) changes by somewhere between -0.025608 and 2.335356 for each 1-year increase in higher education while holding all other predictors (EXP and DENDER) constant.
- With 95% confidence, we estimate that the odds of success (being hired) changes by a factor of somewhere within (e^ -0.025608, e^2.335356) = (0.975, 10.333) for each 1-year increase in higher education while holding all other predictors (EXP and DENDER) constant. This implies that the odds of being hired could decrease by ~ 2.5% or increase by ~ 933.3%, per additional year of education.

9. **Model prediction.** Try out the following code and Describe the difference between the two types. Utilize the R manual pages to understand the use of the function.

```
predict(m, type="link")

predict(m, type="response")
```

```
# Model Predictions

# ?predict
predict(m, type="link")
predict(m, type="response")
```

```
> predict(m, type="link")
         1          2          3          4          5          6          7 
8          9         10         11         12         13         14 
15 
-5.4992779 -4.0250566  3.7437960  1.0142501 -8.7188873 -2.2796685 -2.2053593 
-5.9893414 -6.4091265  4.0892720 -2.2053593 -0.4599712 -7.8090387 -0.9500347 
0.5241866 
        16         17         18         19         20         21         22 
23         24         25         26         27         28 
-3.6795806  0.5944650 -0.8054471 -3.2597955 -3.1152079 -5.0794928 -1.7152958 
5.1437082 -1.4400982 -4.0993657 -0.8054471  5.0734298  3.5992085 
> predict(m, type="response")
           1            2            3            4            5 
6            7            8            9           10           11 
12           13 
0.0040730658 0.0175489472 0.9768829419 0.7338510832 0.0001634423 0.092820865
5 0.0992702542 0.0024990525 0.0016437556 0.9835245625 0.0992702542 0.3869926
554 0.0004058834 
          14           15           16           17           18            1
9           20           21           22           23           24 
25           26 
0.2788778392 0.6281262147 0.0246124941 0.6443889709 0.3088615307 0.036976490
2 0.0424842835 0.0061845776 0.1524780906 0.9941978529 0.1915301359 0.0163126
738 0.3088615307 
          27           28 
0.9937780455 0.9733825060 
```

predict(m, type="link"): The linear predictors of the GLM ($ln\frac{pi}{1-pi}$). The numbers represent the log-odds that the model calculates for each observation based on the fitted model coefficients and the data.

predict(m, type="response"): The predicted probabilities (pi). The linear predictors being transformed back to the probability scale (e^. The numbers represent the model's estimated probabilities of the positive outcome for each observation.

10. **Predict the probability of success of a given case.** Run the following code to predict the probability of being hired for a male candidate who has 6 years of higher education and 3 years of experience.

```
predict(m, newdata=data.frame(EDUC=6, EXP=3, GENDER=1),
          type="response")
```

```
# A male candidate with a master's degree and 6 years of work experience

predict(m, newdata=data.frame(EDUC=6, EXP=3, GENDER=1), type="response")

# Manually calculating the quantity

my.logit <- (-14.2483 + 1.1549 * 6 + 0.9098 * 3 + 5.6037 * 1)
pi <- exp(my.logit) / (1 + exp(my.logit))
pi
```

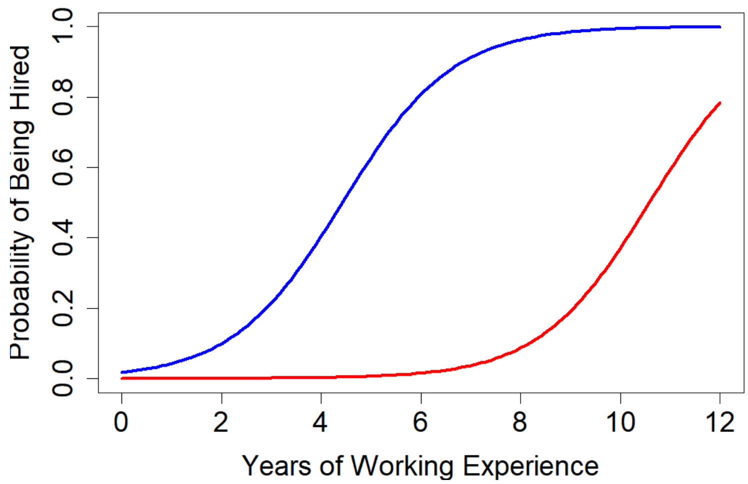
```
> predict(m, newdata=data.frame(EDUC=6, EXP=3, GENDER=1), type="response")
        1 
0.7338511 
> 
> # Manually calculating the quantity
> 
> my.logit <- (-14.2483 + 1.1549 * 6 + 0.9098 * 3 + 5.6037 * 1)
> pi <- exp(my.logit) / (1 + exp(my.logit))
> pi
[1] 0.7338413
```

The predicted probability of being hired is 0.7338413 (~73.38%) for a male with 6 years of education and 3 years of experience, according to the model (m).

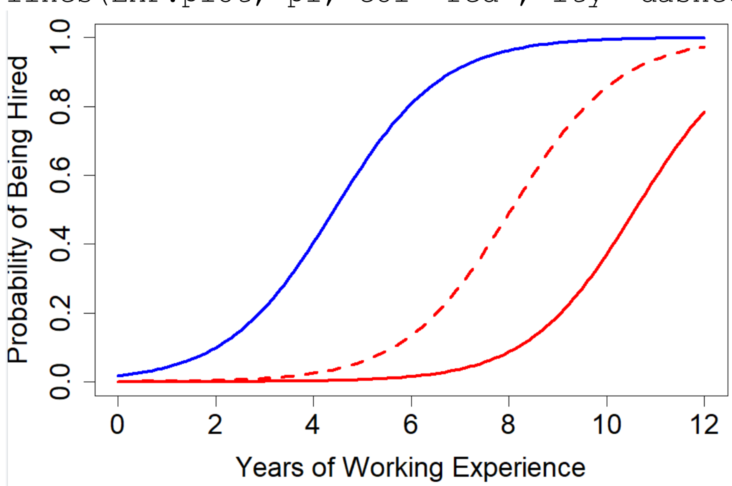
11. **Graphing.**

```
EXP.plot <- seq(0, 12, by=.1)

# GENDER == 0 & EDUC == 4 (Female, Bachelor's degree)

pi <- predict(m, newdata=data.frame(EDUC=4, EXP=EXP.plot,
              GENDER=0), type="response")

plot(EXP.plot, pi, xlim=c(0, 12), ylim=c(0, 1),
     xlab="Years of Working Experience",
     ylab="Probability of Being Hired",
     type='l', col='red', lty="solid", lwd=3,
     cex.axis=1.5, cex.lab=1.5)
```

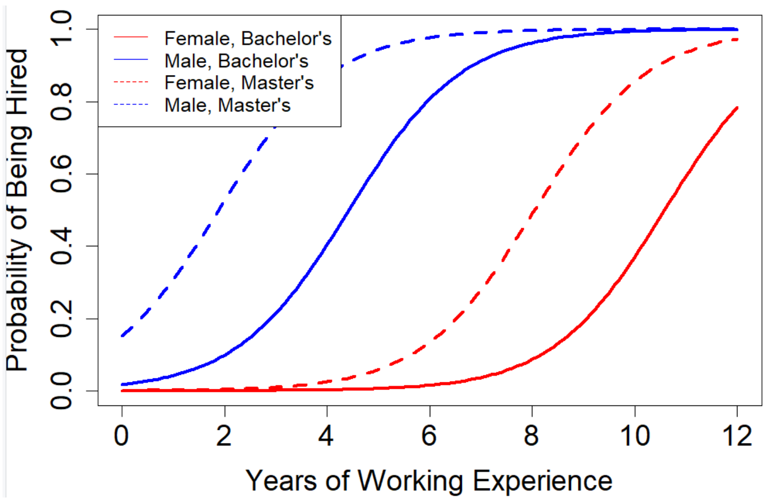

```
# GENDER == 1 & EDUC == 4 (Male, Bachelor's degree)

pi <- predict(m, newdata=data.frame(EDUC=4, EXP=EXP.plot,
              GENDER=1), type="response")

lines(EXP.plot, pi, col='blue', lty="solid", lwd=3)
```




```
# GENDER == 0 & EDUC == 6 (Female, Master's degree)

pi <- predict(m, newdata=data.frame(EDUC=6, EXP=EXP.plot,
              GENDER=0), type="response")

lines(EXP.plot, pi, col='red', lty="dashed", lwd=3)
```




```
# GENDER == 1 & EDUC == 6 (Male, Master's degree)

pi <- predict(m, newdata=data.frame(EDUC=6, EXP=EXP.plot,
              GENDER=1), type="response")
lines(EXP.plot, pi, col='blue', lty="dashed", lwd=3)

legend(x="topleft", legend=c("Female, Bachelor's",
       "Male, Bachelor's", "Female, Master's",
       "Male, Master's"),
       col=c("red", "blue", "red", "blue"),
       lty=c("solid", "solid", "dashed", "dashed"))
```




12. **Derive the <u>operational form</u> of the logistic regression model from the <u>original form</u> (Slide #13).** You can hand-write it on a piece of paper, take a photo of it and insert it here.

$$\mathbf{E\{HIRE\} = Pi} = \frac{\mathbf{exp}(-14.2483 + 1.1549(\mathrm{EDUC}) + 0.9098(\mathrm{EXP}) + 5.6037(\mathrm{GENDER.male})}{1 + \mathbf{exp}(-14.2483 + 1.1549(\mathrm{EDUC}) + 0.9098(\mathrm{EXP}) + 5.6037(\mathrm{GENDER.male})}$$

$$= \frac{1}{1 + \mathbf{exp}(-(-14.2483 + 1.1549(\mathrm{EDUC}) + 0.9098(\mathrm{EXP}) + 5.6037(\mathrm{GENDER.male}))}$$