STP 530: Applied Regression Analysis

Name : **Sai Swaroop Reddy Vennapusa**

Homework **10**

Instructor : **Yi Zheng**

Due Date : 21st Nov 2023, 10:30AM

Steps 1-4:

R code:

```r
# Question 2
install.packages("car")
library(car)

# Question 3
hire.data <- read.csv("~/Downloads/Assignments/STP530/HW10/DISCRIM.csv")

# Question 4
summary(hire.data)
table(hire.data$HIRE)
table(hire.data$GENDER)
pairs(hire.data)
```

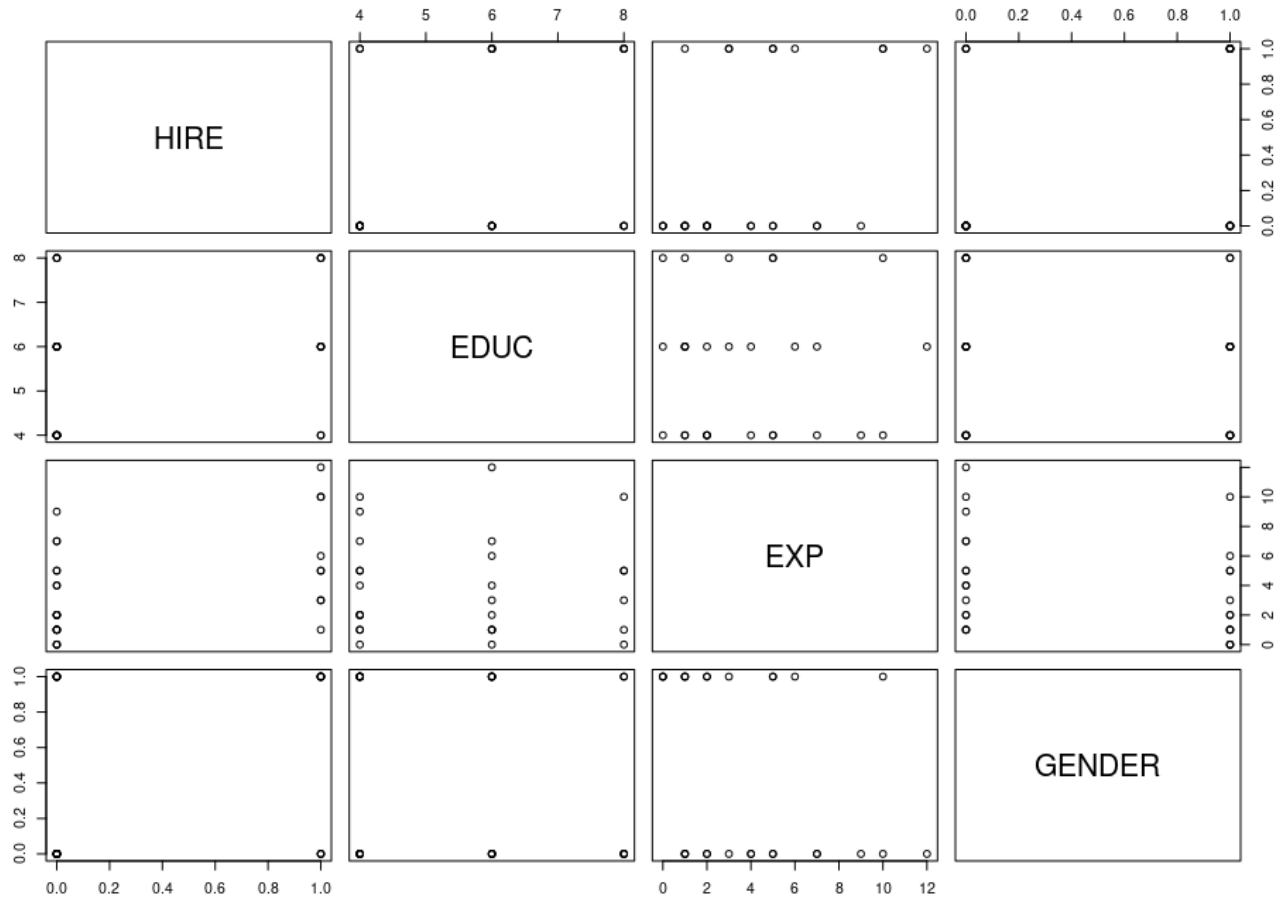
R Output:

```
> library(car)
Loading required package: carData
> # Question 3
> hire.data <- read.csv("~/Downloads/Assignments/STP530/HW10/DISCRIM.csv")
> # Question 4
> 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
```

Output of pairs(hire.data)



Step 5:

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

Output:

```
> 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
```

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

Answer:

The fitted logistic regression model can be written as follows:

ln(P(HIRE=1)/(1-P(HIRE=1))) = -14.2483 + 1.1549 *EDUC + 0.9098 *EXP + 5.6037 *GENDER

Here's what each term represents:

- P(HIRE = 1) is the probability of being hired.
- EDUC is the number of years of higher education.
- EXP is the number of years of work experience.
- GENDER is the gender, where 1 represents male and 0 represents female.

Step 7:
“ 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
b. Work experience
c. Gender ”

Answer:

a. Higher education (EDUC)

Interpretation 1: For every additional year of higher education, the log-odds of being hired increases by 1.1549, assuming that work experience and gender remain constant.

Interpretation 2: Each additional year of higher education is associated with an increase in the odds of being hired by a factor of e^{1.1549}, which is approximately e^{1.1549} is approximately 3.17, holding work experience and gender constant.

Interpretation 3: If the coefficient for higher education is greater than zero, which it is in this case, for every additional year of higher education, the odds of being hired increases by (e^{1.1549} - 1) * 100% is approximately 217%, holding other variables constant.

b. Work experience (EXP)

Interpretation 1: For every additional year of work experience, the log-odds of being hired increases by 0.9098, assuming that higher education and gender remain constant.

Interpretation 2: Each additional year of work experience is associated with an increase in the odds of being hired by a factor of e^{0.9098}, which is approximately e^{0.9098} is approximately 2.48, holding higher education and gender constant.

Interpretation 3: For every additional year of work experience, the odds of being hired increases by (e^{0.9098} - 1) * 100% is approximately 148%, holding other variables constant.

c. Gender (GENDER)

Interpretation 1: Being male (GENDER = 1) is associated with an increase in the log-odds of being hired by 5.6037 compared to being female (GENDER = 0), holding education and work experience constant.

Interpretation 2: Being male is associated with an increase in the odds of being hired by a factor of e^{5.6037}, which is approximately e^{5.6037} is approximately 270.76, compared to being female, holding education and work experience constant.

Interpretation 3: Being male increases the odds of being hired by (e^{5.6037} - 1) * 100% is approximately 26,976%, compared to being female, holding other variables constant.

These interpretations are based on a logistic regression model, which predicts the log-odds of the dependent event (in this case, being hired) as a linear combination of the predictors. The coefficients represent the strength and direction of the association between each predictor and the log-odds of the outcome.

Step 8:

"**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")"
```

Answer:

R Output:

```
> # Step 8
> predict(m, type="link")
         1          2          3          4          5          6          7          8          9         10         11
-5.4992779 -4.0250566  3.7437960  1.0142501 -8.7188873 -2.2796685 -2.2053593 -5.9893414 -6.4091265  4.0892720 -2.2053593
        12         13         14         15         16         17         18         19         20         21         22
-0.4599712 -7.8090387 -0.9500347  0.5241866 -3.6795806  0.5944650 -0.8054471 -3.2597955 -3.1152079 -5.0794928 -1.7152958
        23         24         25         26         27         28
 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
0.0040730658 0.0175489472 0.9768829419 0.7338510832 0.0001634423 0.0928208655 0.0992702542 0.0024990525 0.0016437556
          10           11           12           13           14           15           16           17           18
0.9835245625 0.0992702542 0.3869926554 0.0004058834 0.2788778392 0.6281262147 0.0246124941 0.6443889709 0.3088615307
          19           20           21           22           23           24           25           26           27
0.0369764902 0.0424842835 0.0061845776 0.1524780906 0.9941978529 0.1915301359 0.0163126738 0.3088615307 0.9937780455
          28
0.9733825060
```

The `predict` function in R is used to make predictions based on a model object. The `type` argument specifies the type of prediction that should be returned.

When you use `predict(m, type="link")`, you are asking for the predictions on the scale of the linear predictors (i.e., the log-odds for a logistic regression model). This gives you the value of the linear combination of coefficients and predictor variables, which is the value before it is passed through the logistic function.

For example, a predicted log-odds of -5.4992779 means that, before transforming through the logistic function, the model predicts a very low log-odds of success (in this context, being hired) for the first individual.

On the other hand, `predict(m, type="response")` returns the predictions on the scale of the response variable, which means it gives you the probability that the event of interest occurs. This is achieved by transforming the linear predictors (log-odds) through the logistic function to get values between 0 and 1.

For instance, a predicted probability of 0.0040730658 means that the model predicts a very low probability of being hired for the first individual. The output for the third individual, 0.9768829419, indicates a high probability of being hired according to the model's estimation.

In summary, the difference between `type="link"` and `type="response"` is that the former provides the untransformed linear prediction (log-odds), while the latter provides the predicted probability of the event occurring. If you need to understand the influence of each predictor in the log-odds scale, you would look at `type="link"`. If you are more interested in the predicted probabilities of the outcome, you would use `type="response"`.

Step 9:
"**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")"
```

Answer:

R Output:

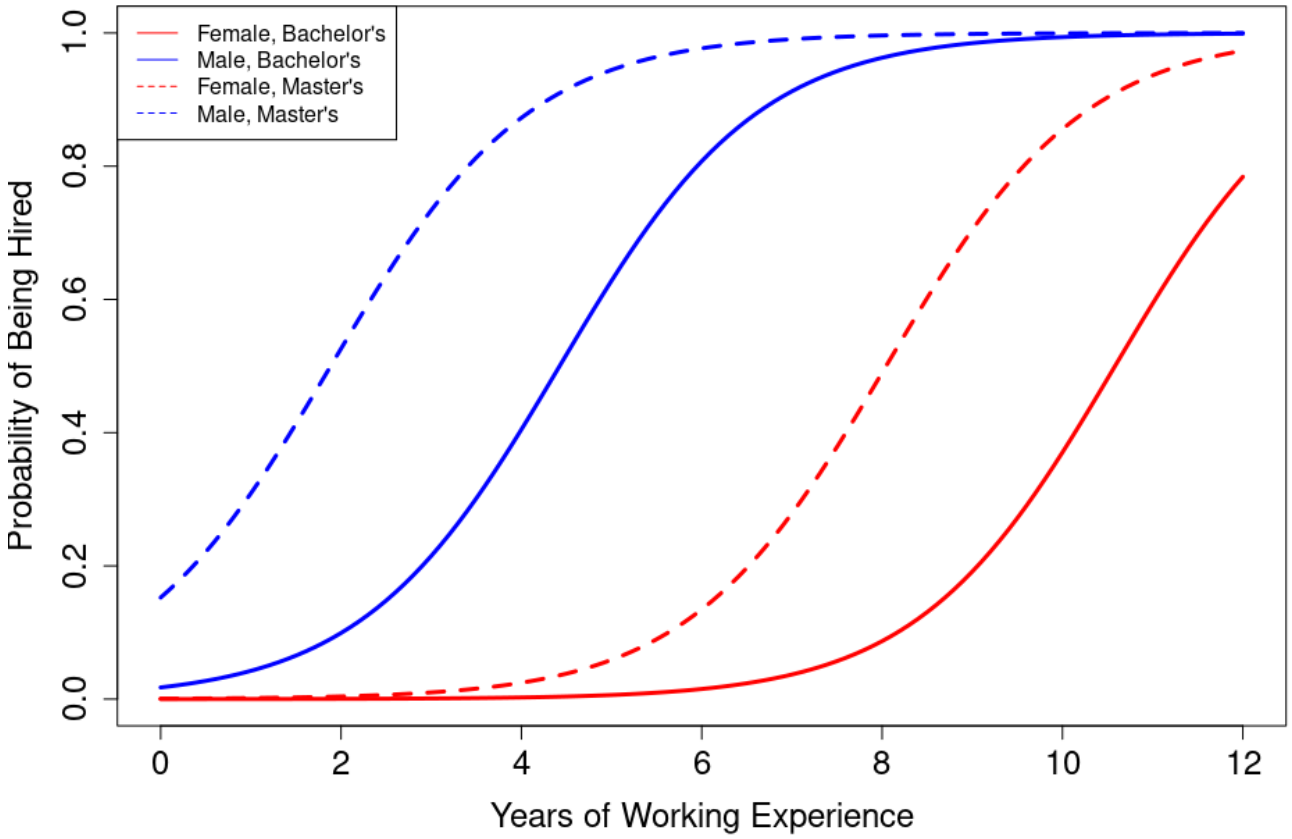
```
> predict(m, newdata=data.frame(EDUC=6, EXP=3, GENDER=1), type="response")
        1
0.7338511
```

Based on the output of the R function `predict` for our logistic regression model, the probability of a male candidate with 6 years of higher education and 3 years of experience being hired is approximately 0.734, or 73.4%. This prediction takes into account the

estimated effects of education, experience, and gender on the likelihood of being hired as indicated by our model coefficients.

Step 10:

"Graphing."



Step 11:

Answer:

The original form of the logistic regression model is given by the logit function:

ln(pi/(1-pi)) = -14.2483 + 1.1549 *EDUC + 0.9098 *EXP + 5.6037 *GENDER

where:
- pi is the probability of being hired.
- beta_0 is the intercept.

- beta_1, beta_2, beta_3 are the coefficients for the predictors EDUC, EXP, and GENDER, respectively.

The estimated coefficients from our model are:
- beta_0 = -14.2483
- beta_1 = 1.1549 for EDUC
- beta_2 = 0.9098 for EXP
- beta_3 = 5.6037 for GENDER

To express the probability pi in terms of the predictors, we take the inverse logit of both sides:

pi = {exp(-14.25 + 1.15(EDU) + 0.91(EXP) + 5.6(GENDER)} / {1 + exp(-14.25 + 1.15(EDU) + 0.91(EXP) + 5.6(GENDER)}

This operational form allows us to calculate the estimated probability of being hired for a candidate with given values of education, experience, and gender.