# STP 530: Applied Regression Analysis

Name : **Sai Swaroop Reddy Vennapusa**

Homework **11**

Instructor : **Yi Zheng**

Due Date : 21st Nov 2023, 11:59PM

**Steps 1-3:**

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
```

**Step 4:**

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 5:**
“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 6:**

**"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. <u>Report and interpret the confidence interval</u>."**

Answer:

To calculate the 95% confidence interval for the slope coefficient of EDUC in the given logistic regression model, we use the formula:

Confidence Interval = b(k) +- z(1-alpha/2).s{b(k)}

Where:
- b(k) is the point estimate of the coefficient.
- s{b(k)} is the standard error of the coefficient.
- z(1-alpha/2) is the z-value for a 95% confidence level.

From the R output, the point estimate b(k) for EDUC is 1.1549, and the standard error s{b(k)} is 0.6023.

For a 95% confidence interval, alpha is 0.05, so 1-alpha/2 is 0.975. The z-value for 0.975 in a standard normal distribution is approximately 1.96.

Now, applying the formula:

Confidence Interval = 1.1549 +- 1.96 * 0.6023

Let's calculate the lower and upper bounds of the confidence interval.

The 95% confidence interval for the slope coefficient of EDUC is approximately (-0.0256, 2.3354).

Interpretation:
- We are 95% confident that the true population parameter for the effect of EDUC on the log-odds of hiring falls within this interval.
- With 95% confidence, we estimate that the log-odds of success (hiring) changes by somewhere between -0.0256 and 2.3354 for each one-unit increase in EDUC, while holding other predictors constant.
- In terms of odds, this means that with 95% confidence, the odds of success change by a factor of somewhere between e^{-0.0256} and e^{2.3354} for each one-unit increase in EDUC, keeping other

variables constant. This indicates that a higher level of EDUC is likely associated with greater odds of hiring, but there is a small probability that EDUC could have a minimal or slightly negative effect.

**Step 7:**
**"Confidence interval of the predicted probability. Use the following code to compute the 95% confidence interval of the predicted probability of being hired for the individual described above. <u>Report and interpret the confidence interval in the context of the problem</u>. "**

Answer:

R Output:

```
> my.pred <- predict(m, newdata=data.frame(EDUC=6, EXP=3,
+                                          GENDER=1), level=.95, type="link", se.fit=T)
> z.crit <- qnorm(p=.975)
> LL.logit <- my.pred$fit - z.crit * my.pred$se.fit
> UL.logit <- my.pred$fit + z.crit * my.pred$se.fit
> LL.pi <- exp(LL.logit) / (1 + exp(LL.logit))
> UL.pi <- exp(UL.logit) / (1 + exp(UL.logit))
> c(LL.pi, UL.pi)
        1         1
0.2680399 0.9540469
```

Interpretation: We are 95% sure that the probability of being hired for a male applicant with 6 years of higher education, 3 years of work experience falls between .268 and .954.

**Step 8:**
**"Likelihood-ratio test of comparing two nested models. The likelihood-ratio test of comparing two nested models is analogous to the general linear test approach F-test. Follow prompts below to complete the 5 steps of the test comparing the *full model*, which we have examined so far, against the *reduced model* where the term GENDER is dropped from the model."**

Answer:

a. Step 1: Assumptions.

- Large sample
- Independent observations

b. Step 2: Hypotheses:
H 0 : beta_3 = 0
H 1 : beta_3 != 0
The full model: logit(HIRE) = beta_0 + beta_1 (EDU) + beta_2 (EXP) + beta_3 (GENDER)
The reduced model: logit(HIRE) = beta_0 + beta_1 (EDU) + beta_2 (EXP)

c. Step 3: Compute the test-statistic.

The test statistic for the likelihood-ratio test is calculated by the difference in deviance between the reduced model and the full model:
G^2 = Deviance{R} - Deviance{F}

G^2 = 26.056 - 14.735 = 11.321

Summary of full model:

```
> 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
```

Summary of reduced model:

```
> summary(m.R)

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

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.4419     2.3361  -2.330   0.0198 *
EDUC          0.5404     0.3287   1.644   0.1002
EXP           0.3717     0.1670   2.226   0.0260 *
---
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: 26.056  on 25  degrees of freedom
AIC: 32.056

Number of Fisher Scoring iterations: 5
```

d. Step 4: Find the p-value.
The p-value is found using the chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the reduced model and the full model. Since the reduced model has 25 degrees of freedom and the full model has 24 degrees of freedom, the degrees of freedom for the test is 1. The p-value is then found by:

p-value = pchisq(q = 11.321, df = 1, {lower.tail=F}) = 0.000766

e. Step 5: Make a conclusion.
Because the p-value is less than the pre-determined significance level $\alpha = 0.05$, reject H 0 and conclude that the full model fits the data significantly better than the reduced model. GENDER contributes significantly to the model on top of EDU and EXP. There is evidence for gender discrimination after accounting for education and work experience.

**Step 9:**
**"Likelihood-ratio test of global model utility. The likelihood-ratio test of *globel model utility* is a likelihood-ratio test comparing the fitted model with the *null model* where all slope coefficients are set to 0. This is analogous to the F-test of global model utility in the ordinary multiple regression models. Follow prompts below to complete the 5 steps of this test."**

Answer:

a. Step 1: Assumptions.
- Large sample
- Independent observations

b. Step 2: Hypotheses:
H 0 : All slopes beta_k = 0
H 1 : at least one slope beta_k != 0
The full model: logit = beta_0 + beta_1 X 1 + beta_2 X 2 + · · ·
The reduced model: logit = beta_0 (the null model)

c. Step 3: Compute the test-statistic:

G^2 = Deviance(Null) − Deviance(Residual)
G^2 = 35.165 – 14.735 = 20.43

d. Step 4: Find the p-value.
p-value = pchisq(q = 20.43, df = 3, lower.tail=F) = 0.0001382406

e. Step 5: Conclusion.
Because the p-value is less than the pre-determined significance level $\alpha = 0.05$, reject H 0 and conclude that the full model fits the data significantly better than the reduced model. At least one of the predictors is useful for predicting the binary outcome.

**Step 10:**
**a."Use the pertinent objects from Problem 14 above to compute McFadden's pseudo R-squared. <u>Report and interpret the results."</u>**

Answer:
McFadden's Pseudo R 2
R^2[MF] = 1 - (-2*ln(L_fitted)/-2*ln(L_null)
= 1 - Residual deviance/Null deviance
= 1 - 14.735/35.165
= 0.58

where L_fitted is the likelihood of the fittedmodel and L_null is the likelihood of the null model where no predictors are involved.

Interpretation:

A McFadden's Pseudo R^2 value of 0.58 suggests that approximately 58% of the variability in the dependent variable is explained by the model. This indicates a relatively good fit between the model and the observed data, as more than half of the variability is accounted for by the predictors of the model. However, it's important to note that Pseudo R^2 values are typically lower than those for regular R^2 and values of 0.2 to 0.4 can be considered as indicative of a strong relationship. Thus, a value of 0.58 would generally be considered quite high for a logistic regression model.

**b. "Run the code below to compute the Tjur's pseudo R-squared. <u>Report and interpret the results</u>. (Note: effectively, only the first and last lines of the code are needed to calculate the Tjur's pseudo R-squared. The other code in between are included to help you understand the process."**

Answer:

R Output:

A Tjur Pseudo R^2 value of 0.621 indicates that there is 62.1% difference in the mean predicted probabilities between the group where the event occurs and the group where it does not. This suggests that the model has good discriminative ability to differentiate between the outcomes. In other words, the model is accurately capturing the difference in probabilities between the two outcome groups, which implies a strong relationship between the predictors and the response variable in the logistic regression model. However, like other Pseudo R^2 measures, this does not have a direct analog to the R^2 in linear regression and shouldn't be interpreted as explaining 62.1% variance in the dependent variable as one would in regular R^2. Instead, it's about the separation of the outcome probabilities predicted by the model.

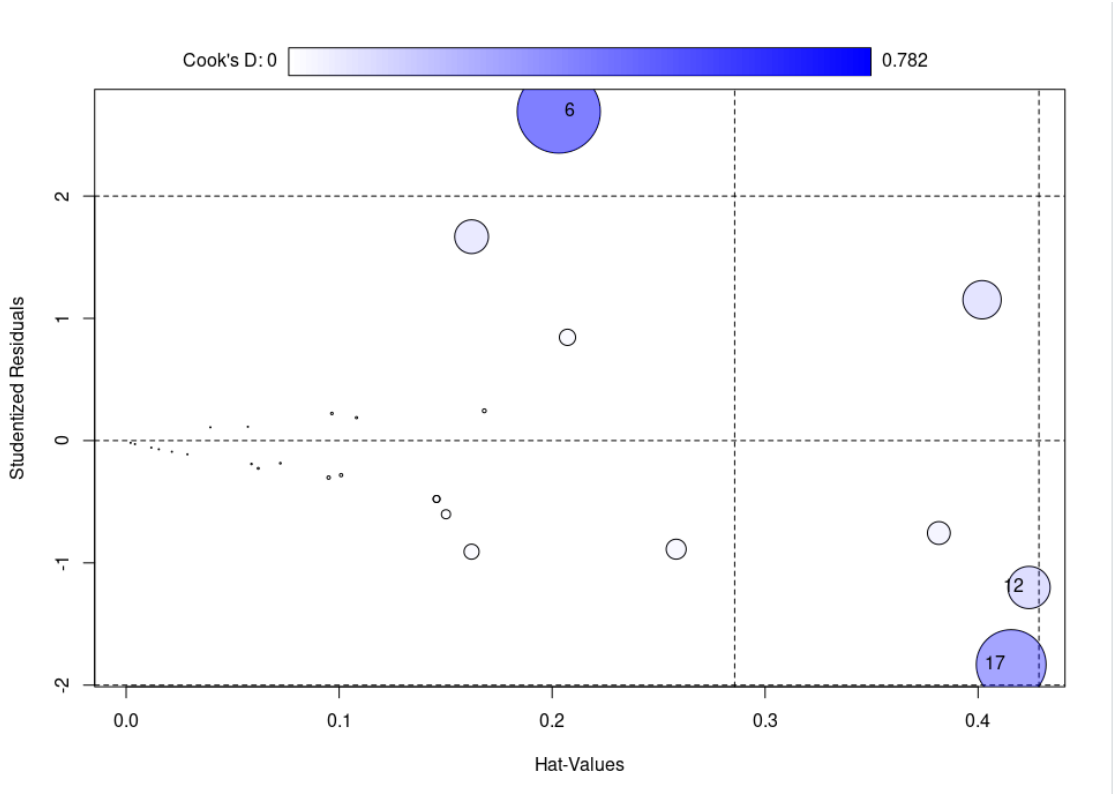
```
> sel <- (hire.data$HIRE == 1)
> lab11data[sel,]
Error: object 'lab11data' not found
> hire.data[sel,]
   HIRE EDUC EXP GENDER
3     1    6   6      1
4     1    6   3      1
6     1    8   3      0
10    1    8  10      0
15    1    4   5      1
18    1    6   1      1
23    1    8   5      1
27    1    4  10      1
28    1    6  12      0
> hire.data[!sel,]
   HIRE EDUC EXP GENDER
1     0    6   2      0
2     0    4   0      1
5     0    4   1      0
7     0    4   2      1
8     0    4   4      0
9     0    6   1      0
11    0    4   2      1
12    0    8   5      0
13    0    4   2      0
14    0    6   7      0
16    0    6   4      0
17    0    8   0      1
19    0    4   7      0
20    0    4   1      1
21    0    4   5      0
22    0    6   0      1
24    0    4   9      0
25    0    8   1      0
26    0    6   1      1
> predict(m, type="response")[sel]
         3          4          6         10         15         18         23         27         28
0.97688294 0.73385108 0.09282087 0.98352456 0.62812621 0.30886153 0.99419785 0.99377805 0.97338251
> mean(predict(m, type="response")[sel])
[1] 0.7428251
> predict(m, type="response")[!sel]
           1            2            5            7            8            9           11           12           13
0.0040730658 0.0175489472 0.0001634423 0.0992702542 0.0024990525 0.0016437556 0.0992702542 0.3869926554 0.0004058834
          14           16           17           19           20           21           22           24           25
0.2788778392 0.0246124941 0.6443889709 0.0369764902 0.0424842835 0.0061845776 0.1524780906 0.1915301359 0.0163126738
          26
0.3088615307
> mean(predict(m, type="response")[!sel])
[1] 0.1218197
> mean(predict(m, type="response")[sel]) - mean(predict(m, type="response")[!sel])
[1] 0.6210054
```

**Step 11: Diagnostics.**

Answer:
a. Influential Points:

R Output:

```
> car::influencePlot(m)
        StudRes       Hat     CookD
6     2.691732 0.2031239 0.7815681
12   -1.201389 0.4239325 0.2016170
17   -1.831995 0.4155453 0.5510990
```




Here are some observations from the influence plot:

1. Observation 6 has a relatively high Cook's distance of approximately 0.782, which suggests it has a substantial influence on the model's coefficients. Its leverage is also high, and it has a large standardized residual, indicating that it does not fit well with the model compared to other observations.

2. Observations 12 and 17, while not as influential as observation 6, still show notable influence on the model. Observation 12 has a moderate Cook's distance of about 0.202, but it has the highest leverage of all points, which means it has a strong influence on the shape of the regression line. Observation 17 has a Cook's distance of approximately 0.551, which is quite substantial, and its leverage is similar to observation 12, making it another point of interest regarding its influence on the model.

3. The plot shows that most other points have low leverage and low standardized residuals, indicating that they fit well within the model and have little influence on the regression coefficients.

Overall, while most data points do not unduly influence the model, observations 6, 12, and 17 are outliers in terms of their influence and could potentially be distorting the model's predictions. It may be worth investigating these points further to understand why they are outliers and to consider whether the model should be adjusted or if these points should be treated differently in the analysis.

b. Cook's d. :

R Output:

```
> p <- 4
> n <- nrow(hire.data)
> pf(q = 0.782, df1 = p, df2 = (n-p))
[1] 0.4520179
```

Interpretation:

The percentile score for the largest Cook's d value of 0.782 is approximately 0.452, or 45.2%. This score is below the 50th percentile on the reference F distribution with p degrees of freedom for the numerator and n - p degrees of freedom for the denominator.

Since this percentile score does not exceed the 50th percentile, it suggests that the observation with Cook's d of 0.782 is not considered overly influential based on this criterion. Generally, a value above the 50th percentile might be a cause for concern, indicating an influential observation. In this case, while the observation may have some influence, it is not to the extent that would typically warrant special attention or action such as removal from the dataset.

c. Dfbetas:

R Output:

```
> dfbetas(m)
```

| | (Intercept) | EDUC | EXP | GENDER |
|---|---|---|---|---|
| 1 | -0.014762642 | 0.0105723414 | 0.0147492413 | 0.0160108535 |
| 2 | -0.053419678 | 0.0504607009 | 0.0481919930 | 0.0350299609 |
| 3 | -0.076346519 | 0.0605645806 | 0.0852187192 | 0.0839297296 |
| 4 | -0.332365006 | 0.2852243687 | 0.3404833421 | 0.4414867329 |
| 5 | -0.001009594 | 0.0008457608 | 0.0009319448 | 0.0009576888 |
| 6 | 0.679876340 | 0.0584955250 | -1.0202362324 | -1.3473394626 |
| 7 | -0.179430514 | 0.1971263632 | 0.1227777744 | 0.0699225111 |
| 8 | -0.010683845 | 0.0093419237 | 0.0086625445 | 0.0099936900 |
| 9 | -0.006984585 | 0.0050480283 | 0.0071580858 | 0.0074587495 |
| 10 | -0.066725970 | 0.0651551597 | 0.0719799305 | 0.0479042440 |
| 11 | -0.179430514 | 0.1971263632 | 0.1227777744 | 0.0699225111 |
| 12 | 0.231241365 | -0.6716223326 | -0.0724276004 | 0.3642845882 |
| 13 | -0.002246936 | 0.0019032923 | 0.0020098824 | 0.0021238718 |
| 14 | -0.104995891 | 0.0259017324 | -0.0878855654 | 0.2386493091 |
| 15 | -0.221838932 | -0.1431031002 | 0.6990131430 | 0.7546822918 |

| | | | | |
|---|---|---|---|---|
| 16 | -0.058709245 | 0.0406002189 | 0.0530055069 | 0.0673303368 |
| 17 | 1.007243500 | -1.4731376500 | -0.1706681243 | -0.8020667146 |
| 18 | 0.171985299 | -0.0615128604 | -0.2979583198 | 0.0753601648 |
| 19 | -0.092280577 | 0.0899263810 | 0.0465433189 | 0.0829985571 |
| 20 | -0.103573170 | 0.1034764731 | 0.0854583029 | 0.0581744110 |
| 21 | -0.022647033 | 0.0202818129 | 0.0168949309 | 0.0210117495 |
| 22 | -0.158653106 | 0.1022575490 | 0.2104754536 | 0.0911178376 |
| 23 | -0.026003798 | 0.0242058099 | 0.0239649414 | 0.0242565052 |
| 24 | -0.287494878 | 0.3646072878 | -0.1135391968 | 0.2282154079 |
| 25 | -0.039209007 | 0.0159091282 | 0.0511813731 | 0.0547658952 |
| 26 | -0.089497037 | 0.0320098215 | 0.1550503841 | -0.0392156276 |
| 27 | -0.023964569 | 0.0146570773 | 0.0329084134 | 0.0295003668 |
| 28 | -0.094292988 | 0.0763361508 | 0.1286556732 | 0.0799699076 |

Observation 6 has the largest absolute value of DFbetas for the slope coefficient of GENDER, with a DFbeta of approximately -1.347. DFbetas measures the difference in each coefficient estimate with and without the influence of the observation. In this case, the negative sign indicates that the inclusion of observation 6 in the model leads to a decrease in the estimate of the GENDER coefficient. Given the substantial magnitude of this DFbeta compared to the others, observation 6 is the most influential concerning the GENDER coefficient and could be considered an outlier or a point with high leverage. It may be worth examining this observation more closely to understand its characteristics and influence on the model.