R Lab 14. LOGISTIC REGRESSION

```
> names(Depr)
[1] "ID"               "Gender"           "Guardian_status"  "Cohesion_score"
[5] "Depression_score" "Diagnosis"
> attach(Depr)
> fix(Data)
> summary(Diagnosis)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0000  0.0000  0.0000  0.1572  0.0000  1.0000    2731
```

 # A lot of missing responses marked as NA. Omit them.

```
> Depr1 = na.omit(Depr)
> attach(Depr1); dim(Depr1)
[1] 458   6
```

 # Now, fit the logistic regression model.

```
> fit = glm( Diagnosis ~ Gender + Guardian_status + Cohesion_score, family = binomial )
> summary(fit)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.00832    0.50478   1.998  0.04577 *
GenderMale      -0.68744    0.28848  -2.383  0.01718 *
Guardian_status -0.74835    0.28602  -2.616  0.00889 **
Cohesion_score  -0.04358    0.01046  -4.167 3.09e-05 ***
```

 # All three variables are significant at 5% level, especially the cohesion score (connection to community).

 # <u>Cross-validation</u>.
 # How well does our model predict within the training data?

```
> Prob = fitted.values(fit)
> summary(Prob)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01958 0.07452 0.12990 0.15720 0.21560 0.57710
```

 # We'll classify a student as having a depression if the probability of that exceeds 0.3.

```
> YesPredict = 1*(Prob > 0.3)       # For all Prob > 0.3, we let YesPredict = 1.
                                    # For all Prob <= 0.3, we let YesPredict = 0.
```

 # Then, create a table of true and predicted responses.

```
> table( Diagnosis, YesPredict )
         YesPredict
Diagnosis   0   1
        0  359  27
        1   48  24
```

 # This is not a perfect result, there are some false positive and false negative diagnoses. Overall, we correctly predicted (359+24)/458 = 83.6% of cases. The *training error rate* is only 16.7%. However, among the students who are really depressed, we correctly diagnosed only 1/3.

Prediction

Let's predict the diagnosis for some particular person, a female who lives with both parents, and has an extremely weak connection with community.

```
> predict( fit, data.frame( Gender="Female", Guardian_status=1, Cohesion_score=26 ))

-0.8730466
```

 # This is the predicted logit. Use the logistic function to convert it into a probability

```
> Y0 = predict( fit, data.frame( Gender="Female", Guardian_status=1, Cohesion_score=26 ))
> P0 = exp(Y0)/(1+exp(Y0))
> P0

0.2946208
```

 # This can also be done by the type option.

```
> predict( fit, data.frame( Gender="Female", Guardian_status=1, Cohesion_score=26 ),
type="response")

0.2946208
```

 # A 29% chance of developing depression! Suppose she has an average community connection instead.

```
> predict( fit, data.frame( Gender="Female", Guardian_status=1, Cohesion_score=52 ),
type="response")

0.1185683
```

 # Only an 11.85% chance now.