Logistic regression

In logistic regression we assume that we sample units from a population. For each unit we have the values of independent variables and the value of the response variable which is binary, 0 and 1 say. The model assumes that the values of the variables are created as independent draws from a distribution $(Y, X_1, \ldots, X_m)$ where we assume that

$$P(Y = 1 | X_1 = x_1, \ldots, X_m = x_m) = \frac{\exp(\beta_0 + \sum_{k=1}^n \beta_k x_k)}{1 + \exp(\beta_0 + \sum_{k=1}^n \beta_k x_k)} ,$$

The parameters $\beta_0, \beta_1, \ldots, \beta_m$ are assumed unknown.

The file `logistic.dat` you have the data. Below is the printout from S-PLUS where the parameters are estimated.

```
*** Import Data ***

Import Successful
File name:   /valjhun/mihael/teaching/istat/verstat/data/logit.dat
Data name:   logit

Number of rows:   1000
Number of columns:   3

Columns:
  Name    Type
1 Col1 numeric
2 Col2 numeric
3 Col3 numeric

*** Generalized Linear Model ***

Call: glm(formula = Col1 ~ Col2 + Col3,
          + family = binomial(link = logit),
          + data = logit, na.action = na.exclude,
          + control = list(epsilon = 0.0001, maxit = 50, trace = F))
```

```
Deviance Residuals:
       Min        1Q    Median       3Q      Max
 -2.061219 -1.047069 0.5614318 1.049224 2.122089

Coefficients:
                 Value Std. Error   t value
(Intercept) 0.05832206 0.06777698 0.8604996
       Col2 0.41705971 0.08130599 5.1295077
       Col3 0.56273870 0.07989416 7.0435527

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 1385.81 on 999 degrees of freedom

Residual Deviance: 1252.243 on 997 degrees of freedom

Number of Fisher Scoring Iterations: 3

Correlation of Coefficients:
     (Intercept)       Col2
Col2  0.0623449
Col3 -0.0199886  -0.3277031

Analysis of Deviance Table

Binomial model

Response: Col1

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                   999   1385.810
Col2  1 79.92338       998   1305.887
Col3  1 53.64399       997   1252.243
```

a. For the given data $y_1, y_2, \ldots, y_n$ in $\mathbf{x}_1, \ldots, \mathbf{x}_n$ write the (conditional) likelihood function.

b. Convince yourself (and me) that the parameters are estimated by maximum likelihood.

c. Compute the Fisher information matrix.

d. Convince yourself that the standard errors given in the printout are obtained from the Fisher information matrix.

e. How would you test the hypothesis $H_0\colon \beta_1 = \beta_2 = 0$?