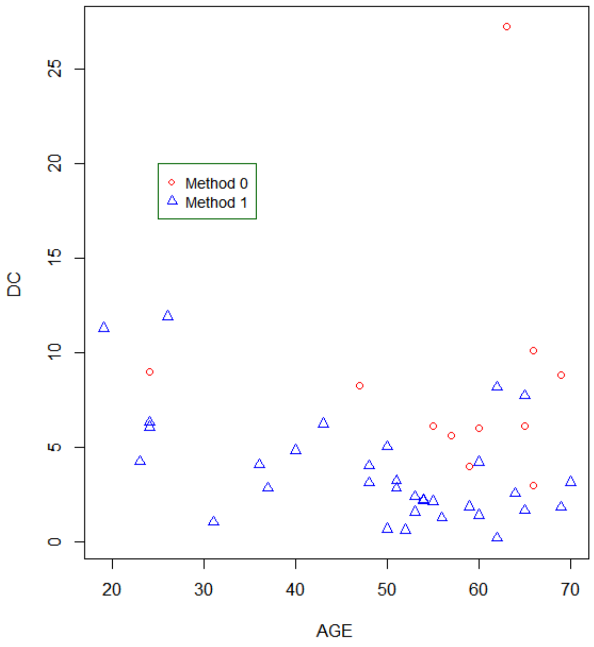
Ex 11.2.1

```
> hw1121 <- read.csv("C:/Users/ma4310/EXR_C11_S02_01.csv")
> x1 <- hw1121$AGE
> x2 <- hw1121$METHOD
> y <- hw1121$DC
> plot(x1, y, xlab="AGE ", ylab="DC ", pch=c(1,2)[as.numeric(x2+1)],col = c("
red","blue")[as.numeric(x2+1)])
> legend(25,20, pch=c(1,2), col=c("red", "blue"), c("Method 0", "Method 1"),
bty="o",  box.col="darkgreen", cex=.9)
```



```
> relation <- lm(y~x1+x2, data = hw1121)
> print(summary(relation))

Call:
lm(formula = y ~ x1 + x2, data = hw1121)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.9503 -1.8686 -0.9299  0.7822 19.0828 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.93349    2.78660   4.641 3.53e-05 ***
x1          -0.07566    0.04388  -1.724 0.092217 .  
x2          -5.47993    1.42945  -3.834 0.000427 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 3.965 on 41 degrees of freedom
Multiple R-squared:  0.2713,   Adjusted R-squared:  0.2358
F-statistic: 7.633 on 2 and 41 DF,  p-value: 0.00152

> confint(relation, level=0.95)
                 2.5 %      97.5 %
(Intercept)  7.3058318 18.56114449
x1          -0.1642726  0.01296244
x2          -8.3667560 -2.59310908
```

Hence we find that the regression line with dummy variables is

$y = 12.93349 - 0.07566\ x_1 - 5.47993\ x_2$

If $x_2=0$, $y = 12.93349 - 0.07566\ x_1$

If $x_2=1$, $y = 7.45356 - 0.07566\ x_1$

Hypothesis testing shows that

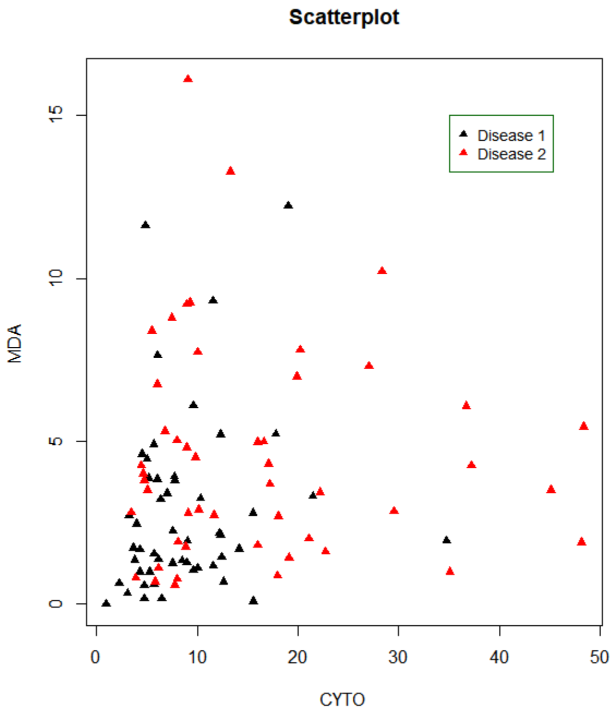
1. **we cannot reject the null hypothesis that $\beta_1=0$ since p-value = $0.092217>0.05$;**
2. **we can reject the null hypothesis that $\beta_2=0$ since p-value = $0.000427<0.05$.**

The 95% confidence interval for β_1 is (-0.1642726 , 0.01296244). Note that the endpoints have different sign, which is consistent with the conclusion of the hypothesis testing.

The 95% confidence interval for β_2 is (-8.3667560, -2.59310908).

Ex 11.2.2

```
> hw1122 <- read.csv("C:/Users/ma4310/EXR_C11_S02_02.csv")
> x1 <- hw1122$CYTO
> x2 <- hw1122$DISEASE
> y <- hw1122$MDA
> # plot different colors
> plot(x1, y, main="Scatterplot", xlab="CYTO ", ylab="MDA ", pch=17, col = fa
ctor(x2))
> legend(35,15, pch=c(17,17), col=c("black", "red"), c("Disease 1", "Disease
2"), bty="o",  box.col="darkgreen", cex=.9)
```



```
> relation <- lm(y~x1+x2, data = hw1122)
> print(summary(relation))

Call:
lm(formula = y ~ x1 + x2, data = hw1122)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8549 -1.8298 -0.8362  1.0222 11.6498 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 1.169398   0.968049   1.208   0.2300  
x1          0.004136   0.033096   0.125   0.9008  
x2          1.621651   0.657308   2.467   0.0154 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 3.057 on 97 degrees of freedom
Multiple R-squared:  0.07011,  Adjusted R-squared:  0.05094
F-statistic: 3.657 on 2 and 97 DF,  p-value: 0.02944

> confint(relation, level=0.95) # CIs for model parameters
                  2.5 %     97.5 %
(Intercept) -0.75191160 3.09070757
x1          -0.06154952 0.06982218
x2           0.31707692 2.92622530
```

Hence we find that the regression line with dummy variables is

$y = 1.169398 + 0.004136\ x_1 + 1.621651\ x_2$

If $x_2=1$, $y = 2.791049 + 0.004136\ x_1$

If $x_2=2$, $y = 4.412700 + 0.004136\ x_1$

Hypothesis testing shows that

- **iii. we cannot reject the null hypothesis that $\beta_1=0$ since p-value = 0.9008>0.05;**
- **iv. we can reject the null hypothesis that $\beta_2=0$ since p-value = 0.0154<0.05.**

The 95% confidence interval for β_1 is (-0.06154952, 0.06982218). Note that the endpoints have different sign, which is consistent with the conclusion of the hypothesis testing.

The 95% confidence interval for β_2 is (0.31707692, 2.92622530).

Ex 11.4.1

```
> hw1141 <- read.csv("C:/Users/EXR_C11_S04_01.csv")
> sex <- hw1141$sex    #X
> vict <- hw1141$vict  #Y
> count <- hw1141$count
> logit_m <- glm(vict~sex, weights=count, family="binomial",  data = hw1141)
> print(summary(logit_m))

Call:
glm(formula = vict ~ sex, family = "binomial", data = hw1141,
    weights = count)

Deviance Residuals:
     1       2       3       4
  8.36  -12.85   11.36  -17.69

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.11918    0.17398  12.180   <2e-16 ***
sex          0.07641    0.21589   0.354    0.723
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 677.12  on 3  degrees of freedom
Residual deviance: 677.00  on 2  degrees of freedom
AIC: 681

Number of Fisher Scoring iterations: 5

> exp(coefficients(logit_m))  # odds ratio
(Intercept)         sex
   8.324324    1.079402
> exp(confint(logit_m, level=0.95)) # condint of odds ratio
                2.5 %    97.5 %
(Intercept) 6.0040530 11.899156
sex         0.7011888  1.638508
```

Hence we find that the logistic regression

$\ln(p/(1-p)) = 2.11918 + 0.07641\, x$

Hypothesis testing shows that we cannot reject the null hypothesis that $\beta_1=0$ since p-value = 0.723>0.05.

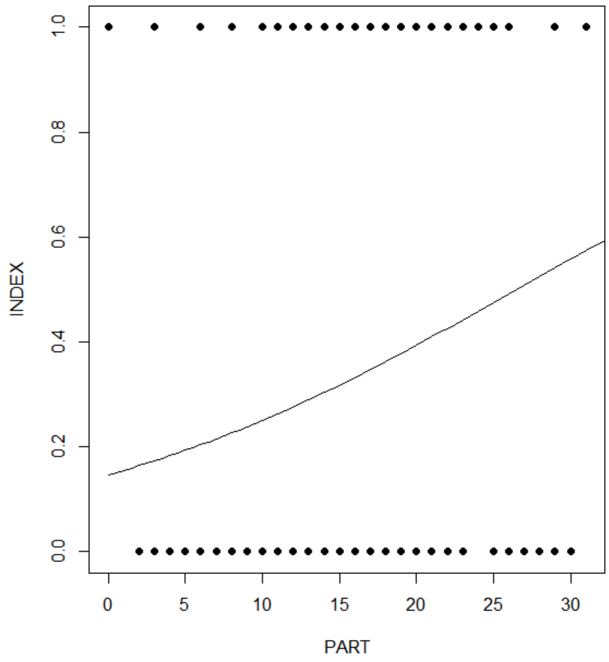
The odds ratio is

$\mathrm{Exp}(\beta_1) = 1.079402$

The 95% confidence interval for $\mathrm{Exp}(\beta_1)$ is (0.7011888, 1.638508). Note that the endpoints does contain the number 1, showing that the odds that a subject being a victim of violence being a m male is not significantly different than the odds that a subject being a victim of violence is a female.

Ex 11.4.2

```
> hw1142 <- read.csv("C:/Users/EXR_C11_S04_02.csv")
> INDEX <- hw1142$INDEX
> PART <- hw1142$PART
> # regression equation
> logit_mod <- glm(PART~INDEX, family="binomial", data = hw1142)
> # plot equation
> plot(INDEX, PART, pch = 16, xlab = "PART", ylab = "INDEX")
> xPART <- seq(0, 45, 0.5)
> yINDEX <- predict(logit_mod, list(INDEX = xPART),type="response")
> lines(xPART, yINDEX)
```



```
> print(summary(logit_mod))

Call:
glm(formula = PART ~ INDEX, family = "binomial", data = hw1142)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.2774  -0.9236  -0.7704   1.2777   1.9591  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.76027    0.45831  -3.841 0.000123 ***
INDEX        0.06641    0.02516   2.639 0.008308 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 236.48  on 183  degrees of freedom
Residual deviance: 229.14  on 182  degrees of freedom
AIC: 233.14

Number of Fisher Scoring iterations: 4

> exp(coefficients(logit_mod)) # model coefficients
(Intercept)       INDEX 
  0.1719981   1.0686667 
> exp(confint(logit_mod, level=0.95))
                 2.5 %    97.5 %
(Intercept) 0.06743088 0.4096688
INDEX       1.01827465 1.1243286
```

Hence we find that the logistic regression

$\ln(p/(1-p)) = -1.76027 + 0.06641\, x_1$

Hypothesis testing shows that we can reject the null hypothesis that $\beta_1=0$ since p-value = 0.008308<0.05.

The odds ratio is

$\text{Exp}(\beta_1) = 1.0686667$

The 95% confidence interval for $\text{Exp}(\beta_1)$ is (1.01827465, 1.1243286). Note that the endpoints do not contain the number 1, showing that the odds that a woman with a high index score will participate are higher that the odds that a woman with a low index score will participate.

QUIZ #7 A study investigated the interrelationship of age of onset of excessive alcohol consumption and cerebrospinal fluid tryptophan (TRYPT) and 5-hidroxyindoleacetic acid (HIAA). We want to use the logistic regression analysis to determine how the odds of finding early onset of excessive alcohol consumption are related to the measurements of TRYPT and HIAA:

(a) Compute and write the multiple logistic regression equation.
(b) Compute the odds ratios.
(c) Test the null hypothesis H0: $\beta_i = 0$ vs H1: $\beta_i \neq 0$ for i=1,2 at significance level 0.05
(d) Compute the 95% confidence interval for beta_1 and beta_2
(e) Compute the McFadden's pseudo r-squared measure

```
> dataq7 <- read.csv("C:/Users/dlabate/Desktop/Teaching/ma4310/dataq7.csv")
> str(dataq7)
'data.frame':	129 obs. of  3 variables:
 $ HIAA : int  57 116 81 78 206 64 123 147 102 93 ...
 $ TRYPT: int  3315 2599 3334 2505 3269 3543 3374 2345 2855 2972 ...
 $ ONSET: int  1 0 1 0 0 1 0 1 1 1 ...
> q7.model = glm(formula = ONSET ~ HIAA+TRYPT, data = dataq7, family = binomi
al)
> summary(q7.model)
Call:
glm(formula = ONSET ~ HIAA + TRYPT, family = binomial, data = dataq7)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9385  -1.3250   0.6931   0.8588   1.4165  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)  
(Intercept)  2.076e+00  1.049e+00   1.979   0.0478 *
HIAA        -1.336e-02  5.512e-03  -2.425   0.0153 *
TRYPT        5.055e-06  2.330e-04   0.022   0.9827  
```

(a) Hence we find that the multiple logistic regression

ln(p/(1-p)) = 2.076 -1.336e-02 **HIAA** + 5.055e-06 **TRYPT**

```
> exp(coefficients(q7.model))
(Intercept)        HIAA       TRYPT 
  7.9695165   0.9867248   1.0000051 
```

(b) The odds ratios are Exp(β_1) = 0.9867248 and **Exp(β_2)** = 1.0000051

(c) Since p-values < 0.05 in the row corresponding to HIAA in the table, we reject the null hypothesis β_1 = 0 (the variable HIAA has a significant role); however, p-values > 0.05 in the row corresponding to TRYP in the table, so we do not reject the null hypothesis β_2 =0

```
> confint(q7.model, level=0.95)
Waiting for profiling to be done...
                    2.5 %        97.5 %
(Intercept)  0.0171940181  4.1671223415
HIAA        -0.0246882900 -0.0028467457
TRYPT       -0.0004330753  0.0004898807
```

(d) The 95% confidence intervals are

95% CI of β_1 `[-0.0246882900 -0.0028467457]`

95% CI of β_2 `[-0.0004330753  0.0004898807]`

```
> library(pscl)
> pR2(q7.model)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU 
-75.82799289 -79.05402563   6.45206550   0.04080795   0.04878581 
0.06905951
```

(e) **The McFadden's pseudo r-squared measure is** 0.04080795