Discrete Analysis Homework 2

110024516 邱繼賢

1.

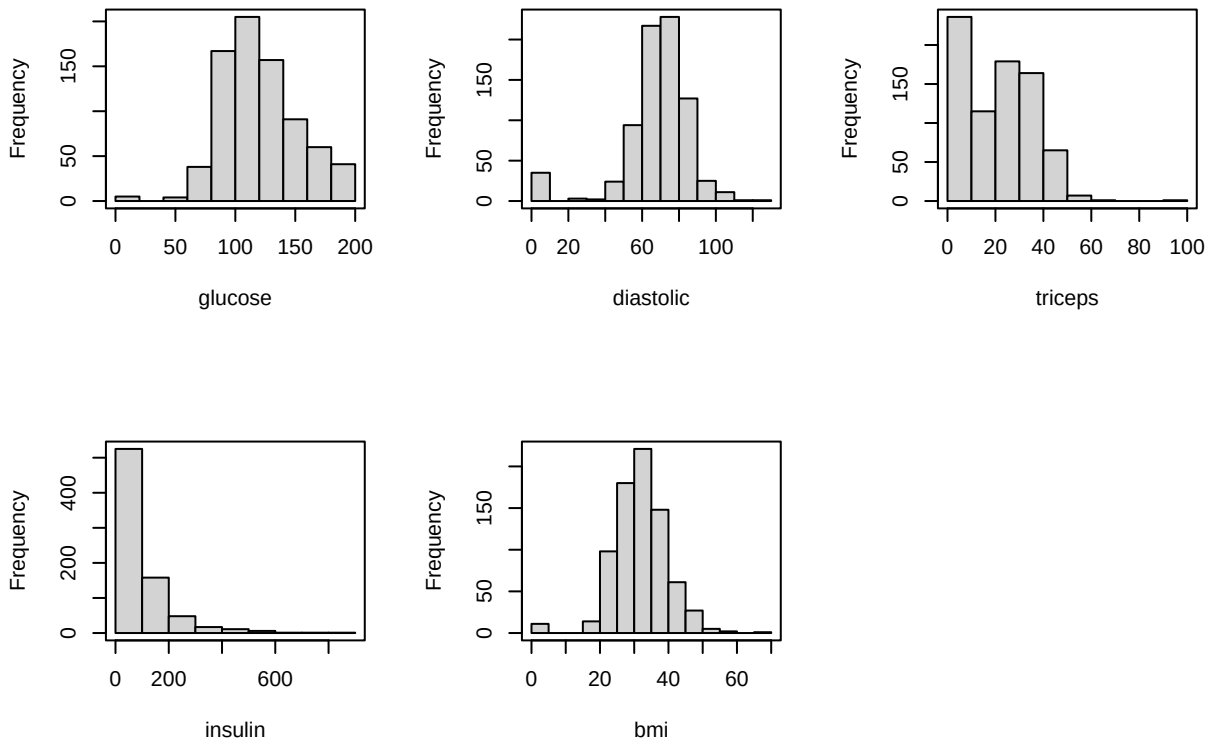
匯入資料後，觀察各變數的數值特徵：

```
library(dplyr)
library(psych)
data = read.table("pima.txt")
data$test = factor(data$test)
summary(data)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     insulin           bmi           diabetes           age        test
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   0:500
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   1:268
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

可發現變數 *glucose, diastolic, triceps, insulin, bmi* 最小值皆為零，不太合理，有可能是紀錄資料者將 missing data 誤植為零，繪製這幾個變數的 histogram 進一步觀察：

```r
par(mfrow = c(2,3))
hist(data$glucose, xlab="glucose", main = "") ; box()
hist(data$diastolic, xlab = "diastolic", main="") ; box()
hist(data$triceps, xlab="triceps", main="") ; box()
hist(data$insulin, xlab="insulin", main="") ; box()
hist(data$bmi, xlab="bmi", main="") ; box()
```



- 變數 *glucose, diastolic, bmi* 皆有一些資料點呈現為零，且遠離於大部分數據的分佈，這很明顯是因為將 missing data 誤植造成的

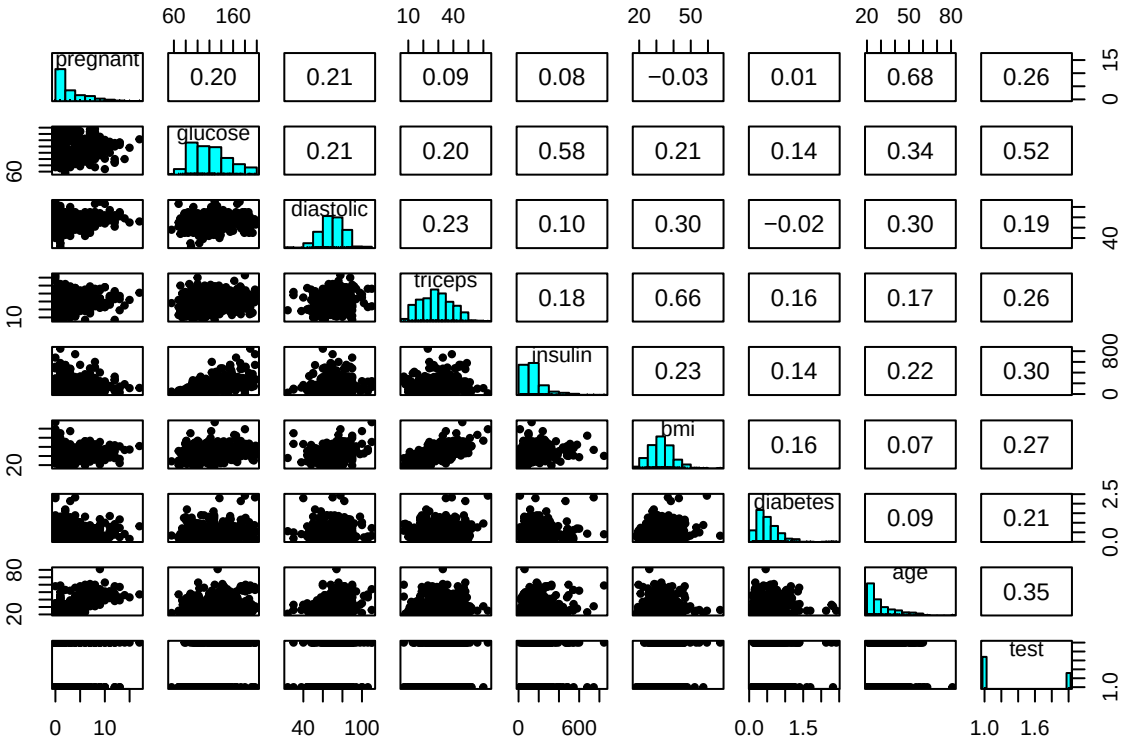
- 變數 *insulin, triceps* 之中數值為零的分佈看起來和大部份數據的分佈是吻合的，不容易判斷這些零值是將 missing data 誤植，還是因為實際測量出的數值太小而記成零

雖然某些零值中可能依舊攜帶著一些資訊，但以我們現在對於資料的理解，並沒有辦法將這些資訊和 missing data 誤植所造成的零值區分開，故選擇將這五個變數中數值為零的資料全部刪除

```r
data2 = data %>%
    filter(glucose*diastolic*insulin*triceps*bmi > 0) %>%
    mutate(test = factor(test))
summary(data2)
```

```
##     pregnant         glucose        diastolic         triceps
##  Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00
##  Median : 2.000   Median :119.0   Median : 70.00   Median :29.00
##  Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15
##  3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00
##  Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00
##     insulin            bmi            diabetes           age        test
##  Min.   : 14.00   Min.   :18.20   Min.   :0.0850   Min.   :21.00   0:262
##  1st Qu.: 76.75   1st Qu.:28.40   1st Qu.:0.2697   1st Qu.:23.00   1:130
##  Median :125.50   Median :33.20   Median :0.4495   Median :27.00
##  Mean   :156.06   Mean   :33.09   Mean   :0.5230   Mean   :30.86
##  3rd Qu.:190.00   3rd Qu.:37.10   3rd Qu.:0.6870   3rd Qu.:36.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200   Max.   :81.00
```

```r
pairs.panels(data2, ellipses = F, smooth = F, density = F)
```



- 資料從原本的 768 的觀測值，減少至 392 個
- 資料中不再呈現不合理為零的數值
- 除了變數 *pregnant, test* 為 discrete(category) 變數，其餘變數皆為 continuous(或視為 approximately continuous) 變數

以除了 *test* 外其餘八個變數的組合當作 covariate classes，然後計算每個組別的糖尿病陽性 (*test=1*) 和陰性 (*test=0*) 人數，將原本的 raw data 轉變成 count data

```
data2_count = data2 %>%
    group_by(pregnant,glucose,diastolic,triceps,insulin,bmi,diabetes,age) %>%
    summarise(positive = sum(test==1), negative = sum(test==0)) %>%
    arrange(desc(positive)) %>%
    mutate(positive=factor(positive),negative = factor(negative))
head(data2_count)
```

```
## # A tibble: 6 x 10
```

```
## # Groups:   pregnant, glucose, diastolic, triceps, insulin, bmi, diabetes [6]
##   pregnant glucose diastolic triceps insulin   bmi diabetes   age positive
##      <int>   <int>     <int>   <int>   <int> <dbl>    <dbl> <int> <fct>
## 1        0      95        85      25      36  37.4    0.247    24 1
## 2        0     104        64      37      64  33.6    0.51     22 1
## 3        0     107        62      30      74  36.6    0.757    25 1
## 4        0     118        84      47     230  45.8    0.551    31 1
## 5        0     121        66      30     165  34.3    0.203    33 1
## 6        0     128        68      19     180  30.5    1.39     25 1
## # ... with 1 more variable: negative <fct>
```

每一組 covariate class 的人數都等於 1，這是一筆 sparse data

2.

建構 generalized linear model：

$$test_x \sim B(1, p_x)$$
$$log\left(\frac{p_x}{1-p_x}\right) = \eta_x = X\beta$$

X 是由 $test$ 以外的其餘八個變數和截距項所形成的 model matrix

```
data2$test = as.numeric(as.character(data2$test))
fit = glm(cbind(test,1-test) ~ pregnant+glucose+diastolic+triceps+insulin+bmi+diabetes+age,
          family = binomial, data = data2)
summary(fit)
```

```
##
## Call:
## glm(formula = cbind(test, 1 - test) ~ pregnant + glucose + diastolic +
##     triceps + insulin + bmi + diabetes + age, family = binomial,
##     data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7823  -0.6603  -0.3642   0.6409   2.5612
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## bmi          7.054e-02  2.734e-02   2.580  0.00989 **
## diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

我們並沒有辦法判斷此模型配飾此筆資料是否適合 (Test for goodness-of-fit)，因為這是一筆每個 covariate class 都只有一個 unit 的 sparse data，所以此模型的 deviance 只為一個 $\hat{p}_x$ 的函數，不包含真實機率 p_x，故我們沒辦法利用 deviance 來判斷模型是否適合。

3.

變數 bmi 的 1st 和 3rd quartile 分別為 $bmi_{1st} = 28.4$ ， $bmi_{3rd} = 37.1$
則此兩個數值的 log odds ratio 為

$$log\left(\frac{O_{3rd}}{O_{1st}}\right) = log(O_{3rd}) - log(O_{1st}) = 37.1\beta_{bmi} - 28.4\beta_{diabetes} = 8.7\beta_{bmi}$$

我們可以用 β_{bmi} 的 MLE 來對其估計，然後取 exp 求出 odds ratio

$$\frac{\hat{O}_{3rd}}{\hat{O}_{1st}} = exp\left(8.7\hat{\beta}_{bmi}\right) = 1.847211$$

```
exp((37.1-28.4)*coef(fit)[7])
```

```
##      bmi
## 1.847211
```

用 profile likelihood method 求出 $\hat{\beta}_{bmi}$ 的 95% 信賴區間

```
library(MASS)
confint(fit)[7,]
```

```
##      2.5 %     97.5 %
## 0.01766988 0.12534312
```

然後對其上下界 $\times 8.7$ 然後取 exp 後即可求出 odds ratio 的信賴區間

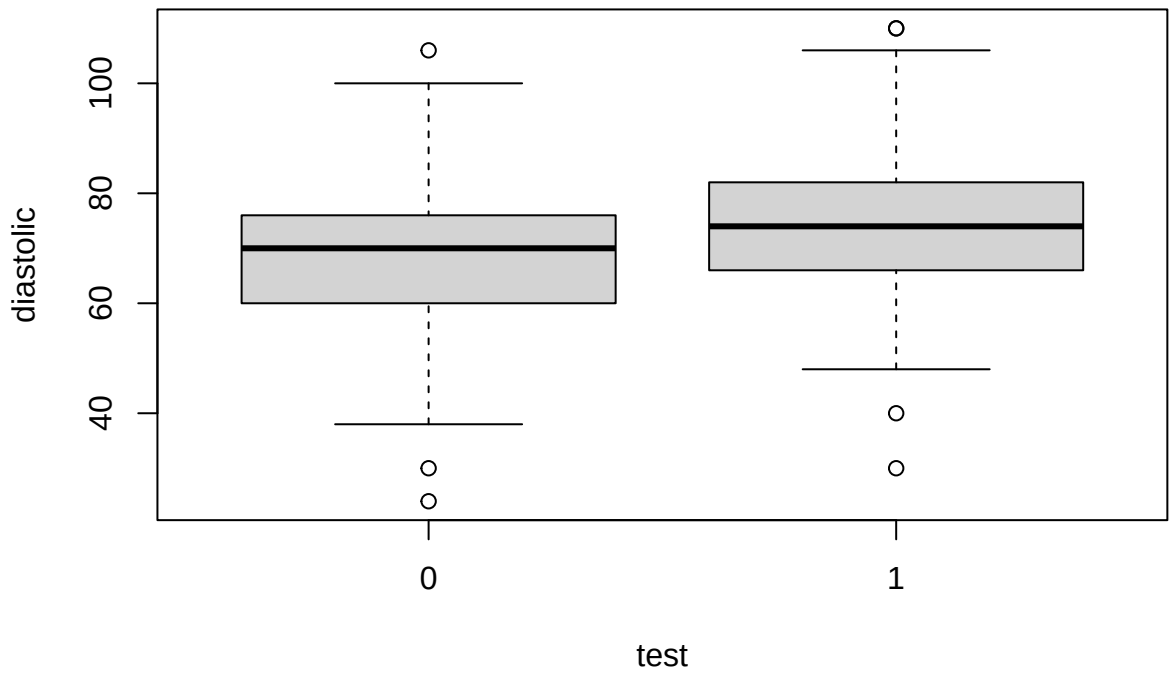
```
exp((8.7)*confint(fit)[7,])
```

```
##    2.5 %   97.5 %
## 1.166174 2.975717
```

4.

(1) Do women who test positive have higher diastolic blood pressures?

```
plot(x=as.factor(data2$test), y=data2$diastolic, xlab = "test", ylab = "diastolic")
```



可以藉由圖形看出糖尿病陽性 (test=1) 時的血壓，整體上高於糖尿病陰性 (test=0) 時的血壓，我們再進一步做檢定確認糖尿病陽性時的血壓平均 μ_1，是否高於糖尿病陰性時的血壓平均 μ_0

$$\begin{cases} H_0 \ : \ \mu_1 \ \leq \ \mu_0 \\ H_1 \ : \ \mu_1 \ > \ \mu_0 \end{cases}$$

```
dias0 = data2$diastolic[data2$test==0]
dias1 = data2$diastolic[data2$test==1]
t.test(dias1,dias0, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  dias1 and dias0
## t = 3.761, df = 237.75, p-value = 0.0001066
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.865015      Inf
## sample estimates:
```

```
## mean of x mean of y
##  74.07692  68.96947
```

p-value $= 0.0001066 < 0.05$，結果呈現顯著，故我們可以推斷出有確診糖尿病的女性血壓較沒確診者高。

(2) Is the diastolic blood pressure significant in the model?

```
drop1(fit, test = "Chi")
```

```
## Single term deletions
##
## Model:
## cbind(test, 1 - test) ~ pregnant + glucose + diastolic + triceps +
##     insulin + bmi + diabetes + age
##           Df Deviance    AIC    LRT  Pr(>Chi)
## <none>         344.02 362.02
## pregnant   1   346.24 362.24  2.214  0.136741
## glucose    1   396.95 412.95 52.929 3.458e-13 ***
## diastolic  1   344.04 360.04  0.014  0.904518
## triceps    1   344.45 360.45  0.431  0.511591
## insulin    1   344.42 360.42  0.397  0.528608
## bmi        1   350.89 366.89  6.871  0.008759 **
## diabetes   1   351.58 367.58  7.559  0.005970 **
## age        1   347.55 363.55  3.529  0.060322 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value $= 0.904518 > 0.05$，結果為不顯著，故變數 *diastolic* 對模型沒有顯著貢獻。

(3) Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

$\hat{\beta}_{diastolic} < 0$ 而且對模型貢獻不顯著，但是根據 **(1)** 我們又知道 *test* 從 0 到 1 時會讓 *diastolic* 的數值連帶跟著上升，這兩個結論看起來是互相矛盾的，我們可以先看回第一題中各變數的相關係數圖表，和第二題中的 summary 報表，會發現

- *diastolic* 和 *glucose, bmi, age* 皆呈現正相關

- *glucose, bmi, age* 的係數 MLE 估計值皆大於零

由此推測可能是因為 *diastolic* 和這三個變數之間具有共線性，影響了 $\beta_{diastolic}$ 的估計值和顯著性，觀察一下 $\hat{\beta}$ 的 estimated covariance matrix $\hat{\Sigma}$

```
summary(fit)$cov.unscaled
```

```
##               (Intercept)      pregnant       glucose     diastolic
## (Intercept)  1.4827307849  1.989842e-03 -3.599164e-03 -5.231847e-03
## pregnant     0.0019898418  3.071991e-03  1.562492e-05 -1.875699e-05
## glucose     -0.0035991637  1.562492e-05  3.326646e-05 -7.509126e-06
## diastolic   -0.0052318469 -1.875699e-05 -7.509126e-06  1.400292e-04
## triceps      0.0018676479 -1.011585e-05  2.847463e-07 -2.071451e-06
## insulin      0.0003606579  3.620494e-06 -3.731717e-06  1.096681e-06
## bmi         -0.0152130926  1.639598e-04  1.647370e-05 -7.928245e-05
## diabetes    -0.1330678793  2.700337e-03  2.421112e-04  2.660592e-04
## age         -0.0048496516 -6.570794e-04 -8.330213e-06 -4.203940e-05
##                   triceps       insulin           bmi      diabetes
## (Intercept)  1.867648e-03  3.606579e-04 -1.521309e-02 -1.330679e-01
## pregnant    -1.011585e-05  3.620494e-06  1.639598e-04  2.700337e-03
## glucose      2.847463e-07 -3.731717e-06  1.647370e-05  2.421112e-04
## diastolic   -2.071451e-06  1.096681e-06 -7.928245e-05  2.660592e-04
## triceps      2.918531e-04  7.718881e-07 -2.770977e-04 -5.023667e-04
## insulin      7.718881e-07  1.706783e-06 -6.413555e-06 -2.666431e-05
## bmi         -2.770977e-04 -6.413555e-06  7.475925e-04  4.628964e-04
## diabetes    -5.023667e-04 -2.666431e-05  4.628964e-04  1.826996e-01
## age         -3.527284e-05 -1.851676e-06  6.241725e-05 -6.157255e-04
##                       age
## (Intercept) -4.849652e-03
## pregnant    -6.570794e-04
## glucose     -8.330213e-06
## diastolic   -4.203940e-05
## triceps     -3.527284e-05
## insulin     -1.851676e-06
## bmi          6.241725e-05
```

```
## diabetes    -6.157255e-04
## age          3.378877e-04
```

可以發現 $\hat{\beta}_{diastolic}$ 對 $\hat{\beta}_{glucose}, \hat{\beta}_{bmi}, \hat{\beta}_{age}$ 的確都呈現負相關，很可能就是這個原因導致 $\hat{\beta}_{diastolic}$ 計算出來後為負值，接下來將變數 *glucose, bmi, age* 從模型中移除再觀察

```
fit2 = update(fit, .~.-glucose-bmi-age)
summary(fit2)
```

```
##
## Call:
## glm(formula = cbind(test, 1 - test) ~ pregnant + diastolic +
##     triceps + insulin + diabetes, family = binomial, data = data2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0265  -0.7837  -0.5225   0.8867   2.1077
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.151343   0.813918  -6.329 2.47e-10 ***
## pregnant     0.148225   0.036930   4.014 5.98e-05 ***
## diastolic    0.019516   0.010362   1.883  0.05965 .
## triceps      0.036866   0.012235   3.013  0.00259 **
## insulin      0.004809   0.001084   4.436 9.16e-06 ***
## diabetes     1.179008   0.359923   3.276  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 410.75  on 386  degrees of freedom
## AIC: 422.75
##
## Number of Fisher Scoring iterations: 4
```

```r
drop1(fit2, test = "Chi")
```

```
## Single term deletions
##
## Model:
## cbind(test, 1 - test) ~ pregnant + diastolic + triceps + insulin +
##     diabetes
##           Df Deviance    AIC     LRT  Pr(>Chi)
## <none>         410.75 422.75
## pregnant   1   427.72 437.72 16.9687 3.800e-05 ***
## diastolic  1   414.37 424.37  3.6211 0.0570507 .
## triceps    1   420.05 430.05  9.2997 0.0022920 **
## insulin    1   433.13 443.13 22.3802 2.237e-06 ***
## diabetes   1   421.95 431.95 11.2002 0.0008179 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

在此情況下，變數 *diastolic* 的係數呈現為正值，並且對模型有顯著貢獻，符合 **(1)** 所得到的結論。