STATISTICS PROJECT

Inference using surrogate outcome data and a validation sample

NILOY RANJAN HORE

STSPCOR18M
Reg. No:- 2014028978
M.Sc Statistics, SEM IV

ACKNOWLEDGEMENT

Firstly, I want to express my sincere gratitude to my project supervisor, ***Dr. Sumanta Adhya***, for his insightful comments and helpful information, practical advice and unceasing ideas that have helped me at all times in my Project. His immense knowledge, profound experience and professional expertise in **surrogate outcome** and **Statistics** have enabled me to complete this project successfully. Without his support and guidance, this project would not have been possible.

I would like to express my sincere gratitude to several individuals and West Bengal State University for supporting me throughout my project.

I am ensuring that this project is finished by me.

Niloy Ranjan Hore
M.Sc Statistics
SEM-IV
West Bengal State University

DECLARATION

*I hereby declare that the project work entitled "**Inference using surrogate outcome data and a validation sample**" submitted in partial fulfilment of the requirements of the M.SC (Statistics) of West Bengal State University is my original work and the project is not submitted previously to any institution for the award of any degree.*

Date:

Signature of the Student:-

:Contents:

1. Abstract
2. Introduction
3. Inference Based on Maximum Likelihood Theory
4. Example
5. Logistic Regression
6. R-code
7. Reference

ABSTRACT:

In the context of estimating β from the regression model $P_\beta(Y|X)$ relating response Y to covariates X, suppose that only a surrogate response S is available for most study subjects. Suppose that for a random subsample of the study cohort, termed the validation sample, the true outcome Y is available in addition to S. We consider maximum likelihood estimation of β from such data and show that it is nonrobust to misspecification of the distribution relating the surrogate to the true outcome, P(S|Y, X). Large-sample distribution theory for maximum estimated likelihood estimates is developed. An illustrative example is presented.

INTRODUCTION:

In many research settings variables of interest are difficult or expensive to obtain. So in many times, a closely related variable may be used as a surrogate for a variable of interest. For example, damage to the heart muscle caused by a myocardial infarction can be assessed accurately using arterioscintography. But, this is an invasive and expensive procedure. Instead, peak cardiac enzyme level in the bloodstream is a more easily obtained variable, and is frequently used as a surrogate measure of heart muscle damage (Wittes, Lakatos & Probstfield, 1989). Also, in school-based smoking prevention projects, current smoking behaviour is generally assessed through self-report using

question-answer. Self-report data are inexpensive to ascertain but may be subject to error. Chemical analysis of saliva samples for the presence of cotinine yields a more accurate measure of current smoking behaviour. However, due to expense, cotinine analysis can only be performed for at most a small subset of subjects enrolled in these large scale studies. Thus in evaluating the results of these studies, self-report of smoking behaviour is used as a surrogate outcome variable. Basically, it is a variable that can be measured (or is easy to measure) that is used in place of one that cannot be measured (or is difficult to measure). For example, whereas it may be difficult to assess the wealth of a household, it is relatively easy to assess the value of a houseThe use of surrogate variables is common particularly in the medical research community and there has been considerable effort to identify 'valid' surrogates. For a review of the use of surrogate variables in clinical trials see a series of papers in Volume 8 of Statistics in Medicine. Little formal methodology exists however for making inference regarding an outcome of interest, Y, when a surrogate outcome variable, S, is used instead. Suppose that the surrogate or auxiliary outcome variable S and covariates X are measured for all study subjects. In addition the true outcome is available for a random subsample of the study cohort termed the validation set V.

INFERENCE BASED ON MAXIMUM LIKELIHOOD THEORY

If we assume a model for P(S|Y, X) which is parameterized by θ and perhaps β then the likelihood for (β, θ) can be expressed as

$$\mathscr{L}(\beta,\theta) = \prod_{i\in V} P_\beta(Y_i|X_i)P_{\beta,\theta}(S_i|Y_i,X_i)\prod_{j\in\bar{V}} P_{\beta,\theta}(S_j|X_j) \quad \ldots 1$$

Here $\bar{V}$ indicates the non validation set, the set of observations on which only the surrogate and covariate data have been measured and

$$P_{\beta,\theta}(S|X) = \int P_\beta(y|X)P_{\beta,\theta}(S|y,X)dy$$

The integral would be replaced with a sum if Y were discrete. Assuming the usual regularity conditions, the maximum likelihood estimates can be found using Newton-Raphson iteration. The numerical calculations are completely straightforward when Y is discrete. If Y is continuous, the contributions to the likelihood score and information by non validation sample observations will generally need to be calculated using numerical integration. In some cases it may be computationally more feasible to find the maximum likelihood estimates using the Expectation-Maximisation (Dempster, Laird & Rubin, 1977) or Multiple-Imputation (Rubin, 1987; Tanner & Wong, 1987) algorithms.
For the case in which the joint distribution of Y and S given X is bivariate normal the calculations are relatively easy. The form of $P_{\beta,\theta}(S|X)$ is also normal with mean $E\{\mu_S(Y,X|X)\}$ and variance $E\{\sigma_s^2(Y,X|X)\} + var\{\mu_s(Y,X|X)\}$ where $\mu_S(Y,X) = E(S|Y,X)$ and $\sigma_S^2(Y,X) = var(S|Y,X)$. For example, if the simple linear regression models hold with $Y = \beta_0 + \beta_1 X + \epsilon_Y$ where ϵ_Y is N(0, σ_Y^2) and independent of X, and with $S = \theta_0 + \theta_1 Y + \theta_2 X + \epsilon_S$ where ϵ_S is $N(0, \sigma_S^2)$ and independent of (Y, X), then a simple linear regression model also holds for the association between S and X, namely,

$$S = (\theta_0 + \theta_1\beta_0) + (\beta_1\theta_1 + \theta_2)X + \epsilon_s^*,$$

Where Engel & Walstra (1991) consider maximum likelihood estimation in the framework of a similar linear model. It is important that the form of the model relating the true and surrogate variables be correctly specified. In the bivariate normal example above, if the true regression model relating S to (Y, X) contained a quadratic term in X, say, but the assumed model omitted X^2, then the maximum likelihood estimate of β_1 based on $\mathscr{L}(\beta,\theta)$ would not be consistent. In a similar vein we performed a small simulation study for binary variables S, Y, X with P(X =1) = 0.5,

$$P(Y|X) = e^{(\beta_0+\beta_1 X)Y}/(1 + e^{\beta_0+\beta_1 X}),$$

Where $\beta_0 = 0, \beta_1 = 0.693$, and

$$P(S|Y,X) = e^{(\theta_0+\theta_1 Y+\theta_2 X)S}/(1 + e^{\theta_0+\theta_1 Y+\theta_2 X})$$

Where $\theta_0 = -0.693, \theta_1 = \theta_2 = 0.693$. In performing the estimation we misspecified P(S|Y, X) by omitting the X term. With sample size n =200 and validation sample size n" = 50 the Newton-Raphson iteration

scheme converged in only 77 of the 100 simulated data sets. Moreover the mean maximum likelihood estimate β_1 in these 77 cases was 0.96 with a standard error of 0.07. This again demonstrates that correct specification of the form of the distribution relating the surrogate and true outcomes is important.

Example:

We constructed the following illustrative example using data from 179 aplastic anaemia patients given bone marrow transplants at the Fred Hutchinson Cancer Research Centre. The objective of this analysis is to relate patient age to incidence of chronic graft versus host disease, GVHD, a complication of the transplant which typically manifests itself after the patient has been discharged from the hospital. Long-term follow-up for these patients is costly. Suppose that chronic GVHD status was ascertained for only 49 of the 179 patients. In fact, we selected this subgroup randomly from the study cohort and ignored chronic GVHD data for the remaining 130 patients. It is known that patients who develop acute GVHD, which manifests itself early in the post-transplant period are at high risk of developing chronic GVHD. We propose to include acute GVHD status as a surrogate for chronic GVHD status for those 130 patients on whom chronic GVHD status is unknown. Patient age X, in years, was dichotomized as X < 20 and $X \geq 20$ at the time of transplant, which separates children from adults essentially. A logistic regression model,

$$P_\beta(Y|X) = e^{(\beta_0+\beta_1 X)Y}/(1+e^{\beta_0+\beta_1 X})$$

was used to relate chronic GVHD (Y) to patient age. A summary of the data is provided in Table 1.

Using only the 49 validation set observations the maximum likelihood estimate for β_1 was found to be 0.74 with a standard error of 0.61. Maximum likelihood estimation based on (1) using all 179 observations

failed to converge. This happened whether a saturated or unsaturated model was used for P(S|Y, X).

Table 1. Numbers of patients by age in years and GVHD status

	Validation set			
Age	Acute GVHD	No chronic GVHD	Chronic GVHD	Non validation set
<20	no	18	4	49
<20	yes	3	6	12
$\geq$ 20	no	8	3	44
$\geq$ 20	yes	1	6	25

Using all 179 observations with the estimated likelihood method, the estimate for β_1 was found to be 0.92 with estimated standard error of 0.62. Therefore, in this example, information provided by the surrogate endpoint did not lead to increased precision in estimating the association between chronic GVHD and patient age using the estimated likelihood method. Since we actually knew the true chronic GVHD status of all 179 individuals we investigated the strength of association between surrogate and chronic GVHD outcomes which would be required for the 130 non validation observations to add usefully to the analysis. With a perfect surrogate, S= observed chronic GVHD, the estimate for β_1 was 1.17 with standard error equal to 0.31. A less than perfect surrogate, but one which was better than observed acute GVHD status used above, was calculated as equal to acute GVHD status with probability 1/2 and equal to chronic GVHD status with probability 1/2. The estimate for was calculated as 1.07 with standard error equal to 0.51. This modest improvement over the analysis using the validation set alone was achieved with a highly predictive surrogate. Indeed for this particular surrogate, in the validation set only 3 of 30 cases with S=0 developed chronic GVHD and 16 of the 19 cases with S=1 developed chronic

GVHD. Whether or not such strength of association between true and auxiliary outcome data will be necessary in general with categorical data and small validation sample fractions is unknown at present.

Logistic regression:

Logistic regression is a method we can use to fit a regression model when the response variable is binary. It is uses a method known as *maximum likelihood estimation* to find an equation of the following form:

$$\log[p(X) / (1-p(X))] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where:

⇒$\mathbf{X_j}$: The j_{th} predictor variable

⇒$\boldsymbol{\beta_j}$: The coefficient estimate for the j_{th} predictor variable

The formula on the right side of the equation predicts the **log odds** of the response variable taking on a value of 1.

For example, we might say that observations with a probability greater than or equal to 0.5 will be classified as “1” and all other observations will be classified as “0”.

Logistic model of data

```
> library(caTools)

> li> split <- sample.split(data, SplitRatio = 0.8)

> train_reg <- subset(data, split == "TRUE")

> test_reg <- subset (data, split == "FALSE")

> logistic_model <- glm(HeartDisease ~
Age+Sex+ChestPainType+RestingBP+Cholesterol+FastingBS+RestingECG+MaxHR+ExerciseAngina+Ol
dpeak+ST_Slope,data = train_reg, family = "binomial")

> logistic_modelbrary(ROCR)


Call:  glm(formula = HeartDisease ~ Age + Sex + ChestPainType + RestingBP +

    Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +

    Oldpeak + ST_Slope, family = "binomial", data = train_reg)


Coefficients:

    (Intercept)             Age            SexM  ChestPainTypeATA  ChestPainTypeNAP   ChestPainTypeTA
RestingBP       Cholesterol

     -1.6925135        0.0084340        1.6083887       -2.2888170       -1.6128705       -1.5677793
0.0061263       -0.0045960

      FastingBS  RestingECGNormal      RestingECGST           MaxHR   ExerciseAnginaY          Oldpeak
ST_SlopeFlat       ST_SlopeUp

      1.0735120       -0.1680743        0.0005634       -0.0028985        0.8799118        0.4851501
1.9141080       -0.5293259


Degrees of Freedom: 688 Total (i.e. Null);  673 Residual

Null Deviance:      946.1

Residual Deviance: 435  AIC: 467

> summary(logistic_model)


Call:

glm(formula = HeartDisease ~ Age + Sex + ChestPainType + RestingBP +

    Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +

    Oldpeak + ST_Slope, family = "binomial", data = train_reg)
```

```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.7835  -0.3697   0.1671   0.4373   2.5804

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.6925135  1.6685339  -1.014 0.310405
Age              0.0084340  0.0159837   0.528 0.597733
SexM             1.6083887  0.3366195   4.778 1.77e-06 ***
ChestPainTypeATA -2.2888170  0.4196082  -5.455 4.91e-08 ***
ChestPainTypeNAP -1.6128705  0.3029734  -5.323 1.02e-07 ***
ChestPainTypeTA  -1.5677793  0.4787022  -3.275 0.001056 **
RestingBP        0.0061263  0.0070702   0.866 0.386216
Cholesterol      -0.0045960  0.0012957  -3.547 0.000390 ***
FastingBS        1.0735120  0.3171896   3.384 0.000713 ***
RestingECGNormal -0.1680743  0.3102862  -0.542 0.588042
RestingECGST     0.0005634  0.4155272   0.001 0.998918
MaxHR            -0.0028985  0.0059727  -0.485 0.627473
ExerciseAnginaY   0.8799118  0.2856626   3.080 0.002068 **
Oldpeak          0.4851501  0.1421184   3.414 0.000641 ***
ST_SlopeFlat     1.9141080  0.4773016   4.010 6.06e-05 ***
ST_SlopeUp       -0.5293259  0.5063192  -1.045 0.295820
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 946.08  on 688  degrees of freedom
Residual deviance: 435.04  on 673  degrees of freedom
AIC: 467.04
```

Number of Fisher Scoring iterations: 6

```
> predict_reg <- predict(logistic_model, test_reg, type = response)
Error in match.arg(type) : object 'response' not found
> predict_reg <- predict(logistic_model, test_reg, type = "response")
>
> predict_reg
        4        11        12        16        23        24        28        35        36        40        47        48
0.834083270 0.013712520 0.842055678 0.118527585 0.007355375 0.775579071 0.028088872
0.009028157 0.025300515 0.789559750 0.223366588 0.036388748
       52        59        60        64        71        72        76        83        84        88        95        96
0.859248644 0.305916983 0.922029729 0.925165948 0.647256823 0.036185063 0.102215044
0.852453059 0.046061445 0.381140082 0.014927390 0.960196676
      100       107       108       112       119       120       124       131       132       136       143
144
0.026276446 0.071259957 0.040446683 0.981262695 0.017459102 0.499967955 0.197281171
0.051935852 0.903829880 0.363933130 0.976958123 0.260586362
      148       155       156       160       167       168       172       179       180       184       191
192
0.006263079 0.026034967 0.988860691 0.095011464 0.992953390 0.037554398 0.056024150
0.068350589 0.071742061 0.940354621 0.319324850 0.060921501
      196       203       204       208       215       216       220       227       228       232       239
240
0.024654778 0.103516982 0.006198097 0.223606152 0.963107276 0.016682881 0.024854377
0.074866179 0.976169839 0.045820207 0.951600000 0.985220804
      244       251       252       256       263       264       268       275       276       280       287
288
0.005792060 0.552838028 0.904125013 0.011467314 0.906357451 0.882314574 0.025854120
0.006218422 0.112761050 0.007612821 0.352346474 0.042998017
      292       299       300       304       311       312       316       323       324       328       335
336
0.010857773 0.972689353 0.892626836 0.879508146 0.680277293 0.784738563 0.489606522
0.628228703 0.977478985 0.926013689 0.665119792 0.860237557
      340       347       348       352       359       360       364       371       372       376       383
384
```

0.610183994 0.928332086 0.969932236 0.789872436 0.829623554 0.691251672 0.931440639 0.816295526 0.990757655 0.069851549 0.984617586 0.644681795

388 395 396 400 407 408 412 419 420 424 431 432

0.982828301 0.996866692 0.969177169 0.944171731 0.990658517 0.987769300 0.989914742 0.807069732 0.791493656 0.953836373 0.993970543 0.350873146

436 443 444 448 455 456 460 467 468 472 479 480

0.796123803 0.994611801 0.983182410 0.823706317 0.985661560 0.886563078 0.512075445 0.958797456 0.024916239 0.993564124 0.984829407 0.969738837

484 491 492 496 503 504 508 515 516 520 527 528

0.919871455 0.814168131 0.965512562 0.760244608 0.977215804 0.587390353 0.134395089 0.555595037 0.973677183 0.529754421 0.960968767 0.037301086

532 539 540 544 551 552 556 563 564 568 575 576

0.987046980 0.989894058 0.342072664 0.932599514 0.939531756 0.097208164 0.760817950 0.959369233 0.986689103 0.940288361 0.990749601 0.979119764

580 587 588 592 599 600 604 611 612 616 623 624

0.988798414 0.962613833 0.538907465 0.308134164 0.804203221 0.971107027 0.764903424 0.634675082 0.145591143 0.929910113 0.950431302 0.873970224

628 635 636 640 647 648 652 659 660 664 671 672

0.069182222 0.263925326 0.979064294 0.044039941 0.643681199 0.011699479 0.770130603 0.087250696 0.887861938 0.577234764 0.006140932 0.990906855

676 683 684 688 695 696 700 707 708 712 719 720

0.288086074 0.017610801 0.134248994 0.052318065 0.044136129 0.937588478 0.958320828 0.051297395 0.969346132 0.816107351 0.960554042 0.904088783

724 731 732 736 743 744 748 755 756 760 767 768

0.542556003 0.062026210 0.339844944 0.711120443 0.188212533 0.057082733 0.245227341 0.487374901 0.946415934 0.041628828 0.251481137 0.011788179

772 779 780 784 791 792 796 803 804 808 815 816

0.966008124 0.979598374 0.234805591 0.710596008 0.853047254 0.984389171 0.273013215 0.485947714 0.479635167 0.021340457 0.453867133 0.089188791

820 827 828 832 839 840 844 851 852 856 863
864

0.942327969 0.885314494 0.140706103 0.015781076 0.837771374 0.125794296 0.069587086
0.838148549 0.078087608 0.949259496 0.022266250 0.850276159

868 875 876 880 887 888 892 899 900 904 911
912

0.208745426 0.400583427 0.007847981 0.074705417 0.067537900 0.969337457 0.159971872
0.031845283 0.318647766 0.042220840 0.037743049 0.978568143

916

0.969697742

#confution matrix

```
> predict_reg <- ifelse(predict_reg > 0.5, 1, 0)

> table(test_reg$HeartDisease, predict_reg)

  predict_reg

    0   1

 0  89  16

 1  14 110

> table(test_reg$HeartDisease, predict_reg)

  predict_reg

    0   1

 0  89  16

 1  14 110

> missing_classerr <- mean(predict_reg != test_reg$HeartDisease)

> print(paste('Accuracy =',1-missing_classerr))

[1] "Accuracy = 0.868995633187773" # accuracy level.

> summary(data)

    Age       Sex     ChestPainType   RestingBP      Cholesterol      FastingBS       RestingECG
MaxHR       ExerciseAngina    Oldpeak

 Min.   :28.00   F:193   ASY:496      Min.   :  0.0   Min.   :  0.0   Min.   :0.0000   LVH   :188   Min.
: 60.0   N:547         Min.   :-2.6000
```

```
 1st Qu.:47.00   M:725   ATA:173      1st Qu.:120.0   1st Qu.:173.2   1st Qu.:0.0000
Normal:552   1st Qu.:120.0   Y:371         1st Qu.: 0.0000

 Median :54.00         NAP:203      Median :130.0   Median :223.0   Median :0.0000   ST    :178
Median :138.0           Median : 0.6000

 Mean   :53.51         TA : 46      Mean   :132.4   Mean   :198.8   Mean   :0.2331            Mean
:136.8              Mean   : 0.8874

 3rd Qu.:60.00                      3rd Qu.:140.0   3rd Qu.:267.0   3rd Qu.:0.0000            3rd
Qu.:156.0              3rd Qu.: 1.5000

 Max.   :77.00                      Max.   :200.0   Max.   :603.0   Max.   :1.0000            Max.   :202.0
Max.   : 6.2000

 ST_Slope    HeartDisease

 Down: 63   Min.   :0.0000

 Flat:460   1st Qu.:0.0000 Up  :395   Median :1.0000  Mean   :0.5534  3rd Qu.:1.0000  Max.
:1.0000
```

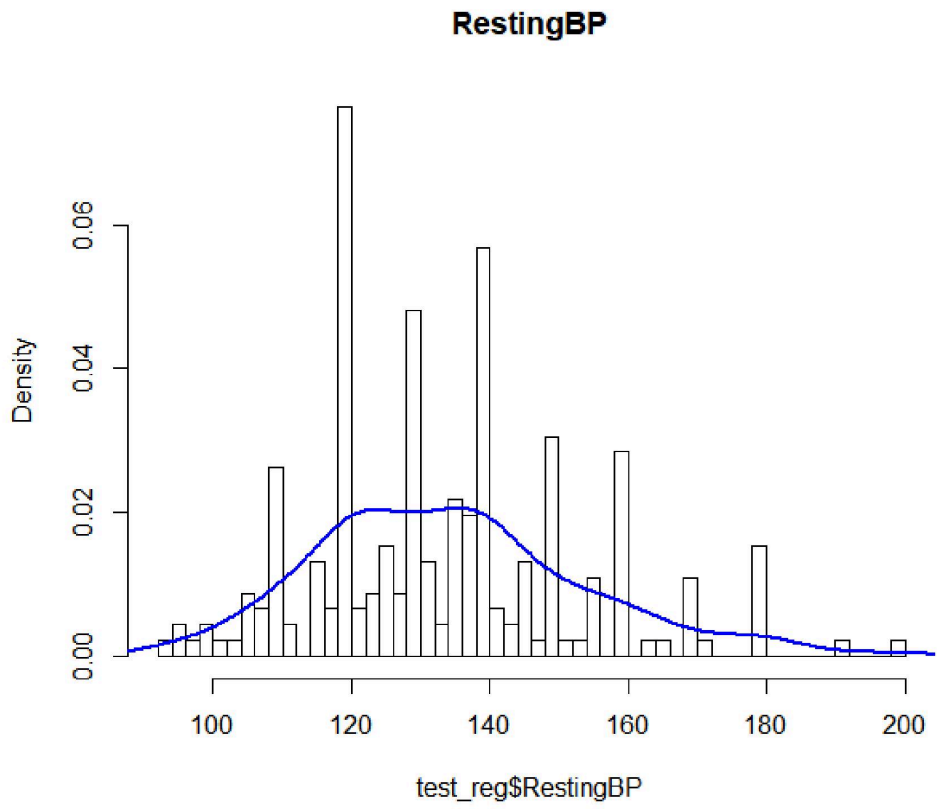
#maximum likelihood estimate of heart disease

```
> hist(test_reg$HeartDisease, breaks = 50, probability = T, main = "HeartDisease")

> lines(density(test_reg$HeartDisease), col = "blue", lwd = 2)

> hist(test_reg$HeartDisease, breaks = 50, probability = T, main = "HeartDisease")

> lines(density(test_reg$HeartDisease), col = "blue", lwd = 2)

> hist(test_reg$RestingBP, breaks = 50, probability = T, main = "RestingBP")

> lines(density(test_reg$RestingBP), col = "blue", lwd = 2)

> nll <- function(theta0, theta1){

+ x <- test_reg$HeartDisease[-idx]

+ y <- test_reg$RestingBP[-idx]

+ mu=exp(theta0 + x*theta1)-sum(y*log(mu)-mu)

> set.seed(200)

> idx <-createDataPartition(test_reg$HeartDisease, p=0.25,list=FALSE)

> est <- stats4::mle(minuslog = nll, start = list(theta0=2,theta1=0))
```

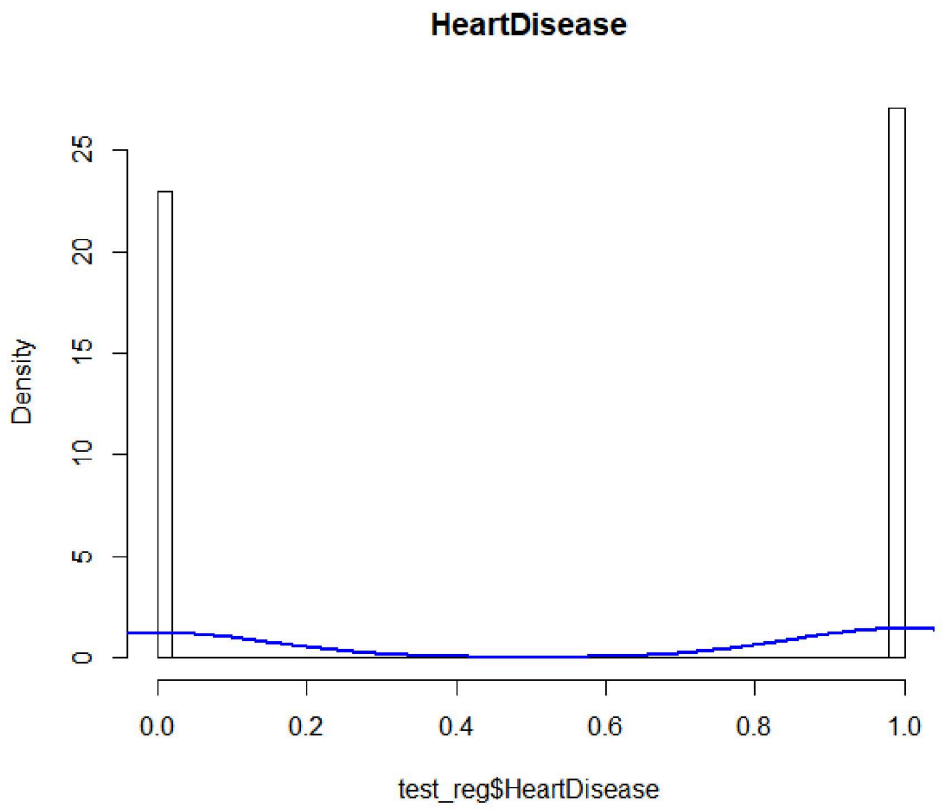
simulation study of parameters

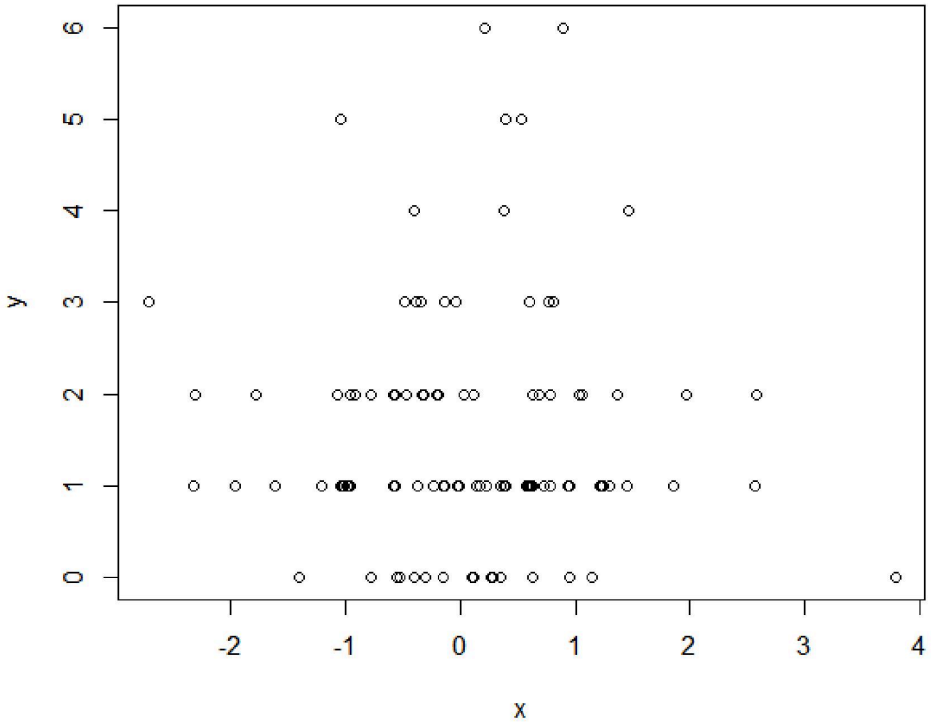
```
> set.seed(200)
```

```
> theta0 <- rnorm(100)
> theta1 <- rnorm(100)
> x <- rnorm(100)
> log.mu <- exp (theta0 + x*theta1)
> y <- r pois (100, exp(log.mu))
> log.mu <- 0.5 +0.3 * theta0
> y <- r pois (100, exp(log.mu))
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    1.00    1.57    2.00    6.00
> plot (x, y)
> log.mu <- 0.5 +0.3 * theta1
> y <- r pois (100, exp(log.mu))
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    1.00    1.00    1.71    2.25    5.00
> plot (x, y)
```

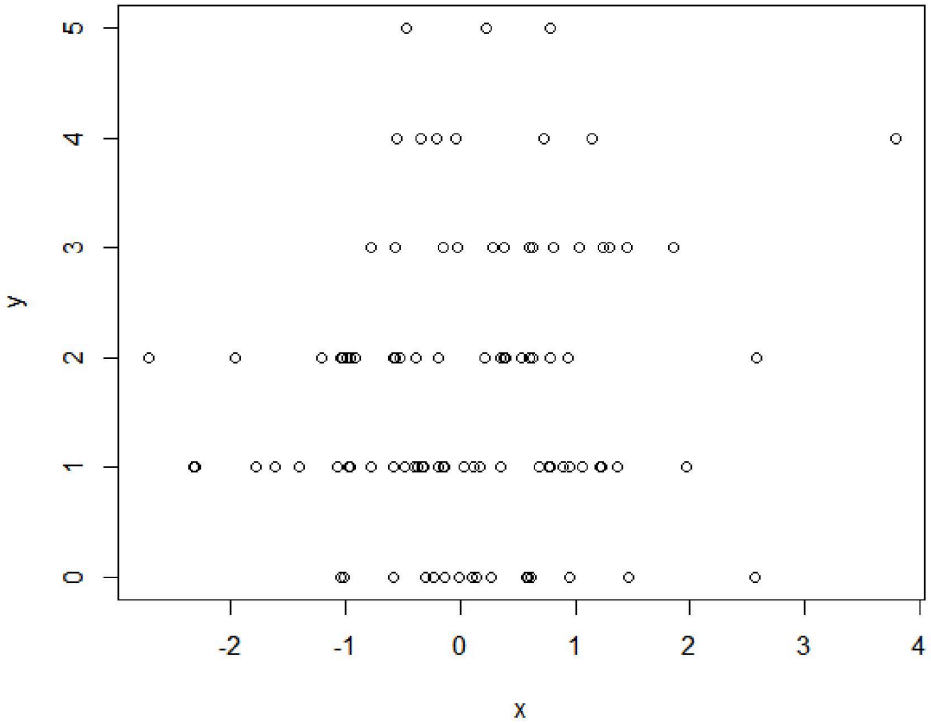

RestingBP
Density
0.00
0.02
0.04
0.06
100
120
140
160
180
200
test_reg$RestingBP

>


HeartDisease
Density
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8
1.0
test_reg$HeartDisease



simulation for theta0



simulation for theta1

Reference:-

1. MARGARET SULLIVAN PEPE, Fred Hutchinson Cancer Research Centre, 1124 Columbia Street, Seattle, Washington 98104, U.S.A. Biometrika (1992). Link:- https://academic.oup.com/biomet
2. WITTES, J., LAKATOS, E. & PROBSTFIELD, J. (1989). Surrogate endpoints in clinical trials: cardiovascular diseases. Statist. Med. 8, 415-25
3. For R-code, using Statology. Link:-Statology