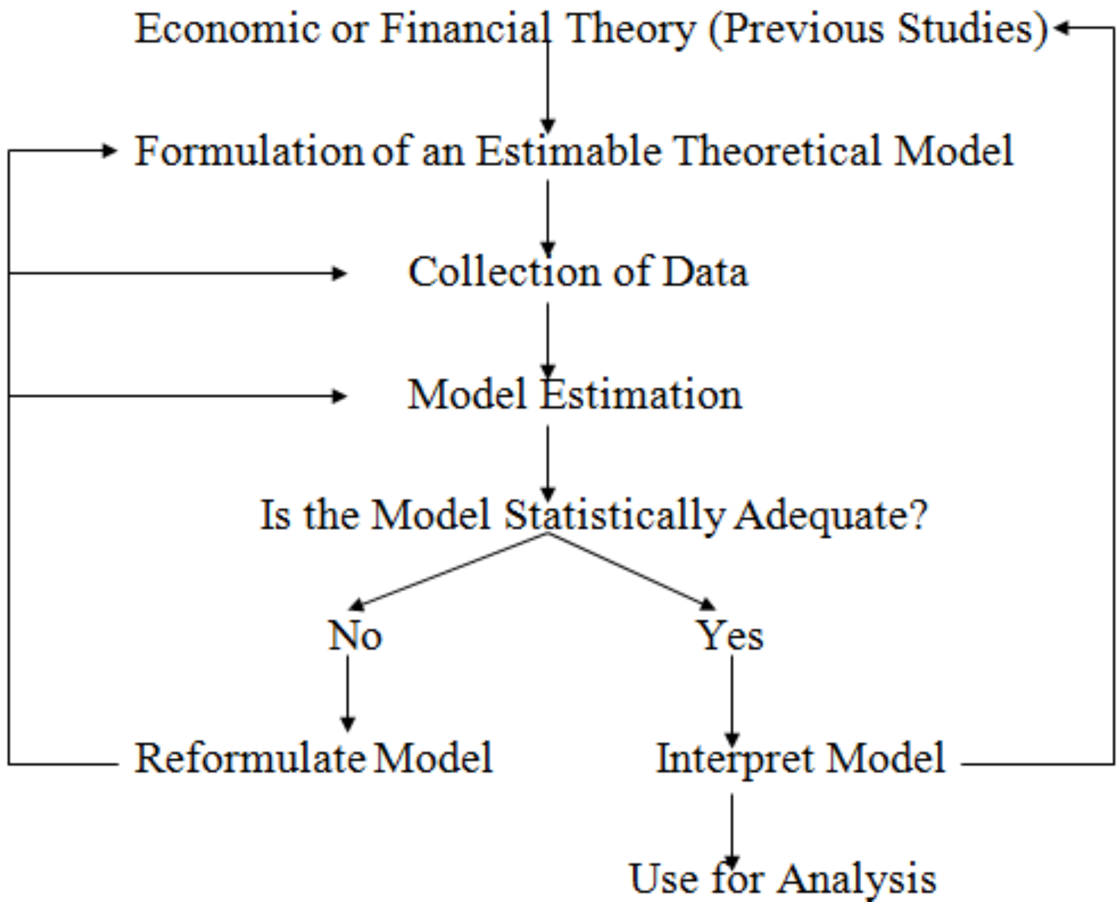
Quantitative Analysis: Time Series Data Modeling and Forecasting - Case Study

Jerry

Case Study

- 某建設公司暨房地產商收集近年的房地產銷售資料以及房地產廣告媒體投放資料，並打算編列去年年度預算的 150% 作為明年度的行銷預算。但是，礙於傳統產業對數據處理生疏，該公司遲遲無法決定應該如何將行銷預算做適當分配。您身為顧問，與公司負責主管進行會談後，發現該公司的過去決策模式偏好使用經驗法則。雖然對於風險與報酬率稍有概念，但是缺乏量化分析以及建立預測模型的能力，並且公司內部也沒有足夠的軟體開發人才。試以量化分析角度結合風險與報酬率概念，為該公司提出最佳預算配置模型。

Steps involved in formulating a forecasting model



General statement of the problem

- 如何在有限預算下，提出 ”最佳” 行銷預算分配？
- This will usually involve the formulation of a theoretical model, or intuition from some theories that two or more variables should be related to one another in a certain way.
- The model is unlikely to be able to completely capture every relevant real-world phenomenon, but it should present a suffi- ciently good approximation that it is useful for the purpose at hand.

General statement of the problem

- 如何在有限預算下，提出 ”最佳” 行銷預算分配？
- 定義何謂最佳
  - 最高的銷售金額？
  - 最高的銷售金額成長率？
  - 最小的銷售風險 (成長率標準差)？

General statement of the problem

- It is preferable not to work directly with asset prices, so we usually convert the raw prices into a series of returns.

Simple returns or log returns

$$R_t = \frac{p_t - p_{t-1}}{p_{t-1}} \times 100\% \qquad\qquad R_t = \ln\left(\frac{p_t}{p_{t-1}}\right) \times 100\%$$

where, R_t denotes the return at time t

p_t denotes the asset price at time t

ln denotes the natural logarithm

General statement of the problem

- Risk aversion assumption (風險趨避假設)
- Utility Function (效用函數)
- Risk-adjusted return (風險調整後收益)
    - E (Return Rate) / Stdev (Return Rate)
- Efficient Frontier (效率前緣)

General statement of the problem

- Risk Aversion and Utility Values
    - Risk averse investors reject investment portfolios that are fair games or worse.
    - These investors are willing to consider only risk-free or speculative prospects with positive risk premiums risk premiums.

General statement of the problem

- Utility Function

$$U = E(r) - \frac{1}{2} A\sigma^2$$

  - U = utility.

    E(r) = expected return on the asset/portfolio.

    A = coefficient of risk aversion. A coefficient of risk aversion.

    σ^2 = variance of returns.

General statement of the problem

- Risk Aversion and Utility Values

Investor Risk Aversion (*A*)	Utility Score of Portfolio *L* [$E(r) = .07$; $\sigma = .05$]	Utility Score of Portfolio *M* [$E(r) = .09$; $\sigma = .10$]	Utility Score of Portfolio *H* [$E(r) = .13$; $\sigma = .20$]
2.0	$.07 - ½ \times 2 \times .05^2 = .0675$	$.09 - ½ \times 2 \times .1^2 = .0800$	$\mathbf{.13 - ½ \times 2 \times .2^2 = .09}$
3.5	$.07 - ½ \times 3.5 \times .05^2 = .0656$	$\mathbf{.09 - ½ \times 3.5 \times .1^2 = .0725}$	$.13 - ½ \times 3.5 \times .2^2 = .06$
5.0	$.07 - ½ \times 5 \times .05^2 = .0638$	$\mathbf{.09 - ½ \times 5 \times .1^2 = .0650}$	$.13 - ½ \times 5 \times .2^2 = .03$

General statement of the problem

- The Indifference Curve

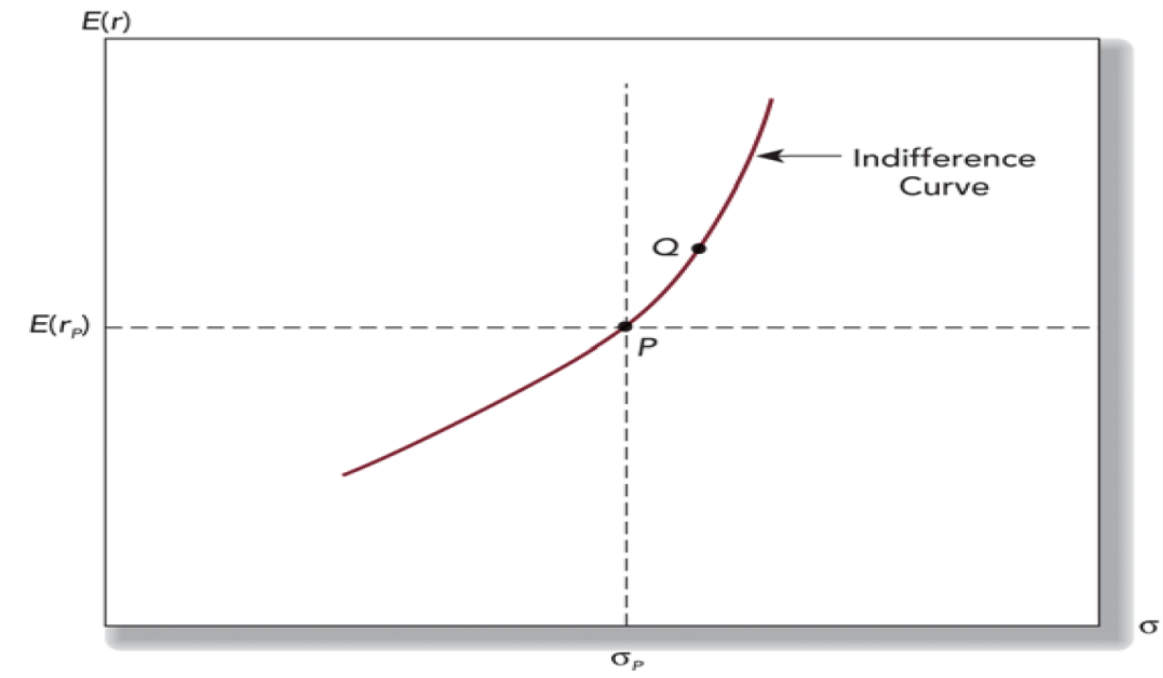
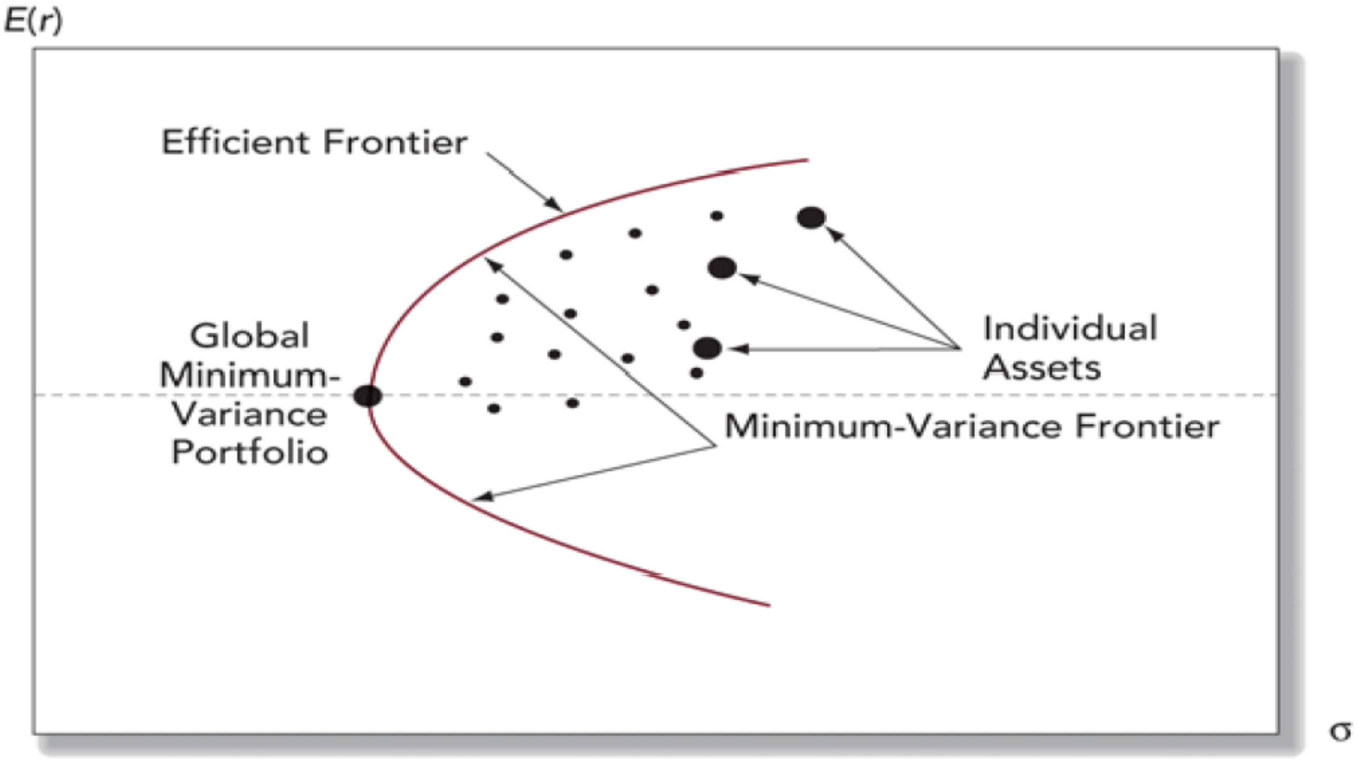


FIGURE 6.2 The indifference curve

General statement of the problem

- The Efficient Frontier



General statement of the problem

- 如何在有限預算下，提出 ”最佳” 行銷預算分配？
  - 最高的銷售金額？
  - 最高的銷售金額成長率？
  - 最高的風險調整後成長率？
  - 最高的效用函數？

Collection of data relevant to the model

- 時間序列資料（Analysis.txt）（對照.txt）
    - 投放費用
    - 客戶參訪
    - 成交量
- Budget 也可以放入變數。
- Time Lag Effect
- 除了以上變數還有什麼需要考慮？
    - Open data: 房地產指數、實價登錄、REIT

Selection of the model

- Regression Analysis
  - Regression is probably the single most important tool
  - It is concerned with describing and evaluating the relationship between a given variable (usually called the dependent variable) and one or more other variables (usually known as the independent variable(s)).

Selection of the model

- Denote the dependent variable by y and the independent variable(s) by x_1, x_2, ... , x_k where there are k independent variables.

- Some alternative names for the y and x variables:

y	x
dependent variable	independent variables
regressand	regressors
effect variable	causal variables
explained variable	explanatory variable

- Note that there can be many x variables.

Selection of the model

- Regression is different from Correlation
  - If we say y and x are correlated, it means that we are treating y and x in a completely symmetrical way.
  - In regression, we treat the dependent variable (y) and the independent variable(s) (x's) very differently. The y variable is assumed to be random or "stochastic" in some way, i.e. to have a probability distribution. The x variables are, however, assumed to have fixed ("non-stochastic") values in repeated samples.

Selection of the model

- We can use the general equation for a straight line,

  $$y=a+bx$$

  to get the line that best "fits" the data.

- However, this equation ($y=a+bx$) is completely deterministic.

- Is this realistic? No. So what we do is to add a random disturbance term, u into the equation.

  $$y_t = \alpha + \beta x_t + u_t$$

  where $t = 1,2,3,4,5,....$

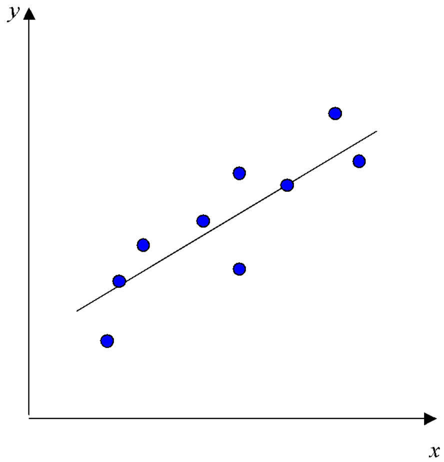
Selection of the model

- Why do we include a Disturbance term?
- The disturbance term can capture a number of features:

  - We always leave out some determinants of y_t
  - There may be errors in the measurement of y_t that cannot be modelled.
  - Random outside influences on y_t which we cannot model

Selection of the model

- Determining the Regression Coefficients
  - So how do we determine what α and β are?
  - Choose α and β so that the (vertical) distances from the data points to the fitted lines are minimised (so that the line fits the data as closely as possible):



Selection of the model

- Determining the Regression Coefficients
  - Ordinary Least Squares (OLS)
  - Maximum Likelihood Estimation (MLE)

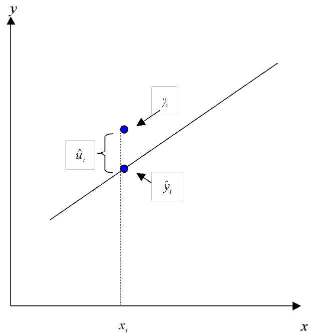
Selection of the model

- So min. $\hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \hat{u}_4^2 + \hat{u}_5^2$, or minimise $\sum_{t=1}^{5} \hat{u}_t^2$. This is known as the residual sum of squares.

- But what was $\hat{u}_t$? It was the difference between the actual point and the line, y_t - $\hat{y}_t$.

- So minimising $\sum (y_t - \hat{y}_t)^2$ is equivalent to minimising $\sum \hat{u}_t^2$ with respect to α and β .

(其實交給套件來跑即可 -> SciPy for Python; R)



Selection of the model

- But what if our dependent (y) variable depends on more than one independent variable?

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \ldots + \beta_k x_{kt} + u_t$$

Selection of the model

- Linearity

  In order to use OLS, we need a model which is linear in the parameters (α and β). It does not necessarily have to be linear in the variables (*y* and *x*).

  Linear in the parameters means that the parameters are not multiplied together, divided, squared or cubed etc.

  Some models can be transformed to linear ones by a suitable substitution or manipulation, e.g. the exponential regression model

$$Y_t = e^{\alpha} X_t^{\beta} e^{u_t} \Leftrightarrow \ln Y_t = \alpha + \beta \ln X_t + u_t$$

  Then let y_t=ln Y_t and x_t=ln X_t

$$y_t = \alpha + \beta x_t + u_t$$

Statistical evaluation of the model

- Statistical Inference
- R squared
- Area under ROC curve
- Test for Pearson Correlation
- Pregibon Test for Linearity
- Ramsey Regression Equation Specification Error Test (RESET)
- Training Group / Validation Group
- ……

Statistical evaluation of the model

- Statistical Inference

We want to make inferences about the likely population values from the regression parameters.

Example: Suppose we have the following regression results:

$$\hat{y}_t = 20.3 + 0.5091 x_t$$
$$(14.38) \quad (0.2561)$$

"β = *0.5091*" is a single (point) estimate of the unknown population parameter, β. How "reliable" is this estimate?

The reliability of the point estimate is measured by the coefficient's standard error.

Statistical evaluation of the model

- Statistical Inference
  - *Distribution of coefficient~ $T_{n-2}(\beta, s.e.(\beta))$*

H_0:	$\beta_1 = 0$	(no linear relationship)
H_1:	$\beta_1 \neq 0$	(linear relationship does exist)

$$T_{n-2} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

Statistical evaluation of the model

- R squared
  - We would like some measure of how well our regression model actually fits the data.
  - We have goodness of fit statistics to test this: i.e. how well the sample regression function (srf) fits the data.
  - The most common goodness of fit statistic is known as R^2.
  - what we are interested in doing is explaining the variability of y about its mean value, $\bar{y}$, i.e. the total sum of squares, *TSS*:

$$TSS = \sum (y_t - \bar{y})^2$$

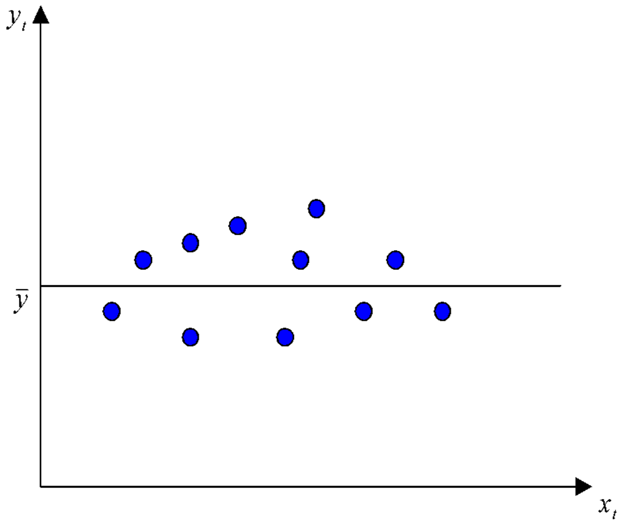
  - We can split the *TSS* into two parts, the part which we have explained (known as the explained sum of squares, *ESS*) and the part which we did not explain using the model (the *RSS*)

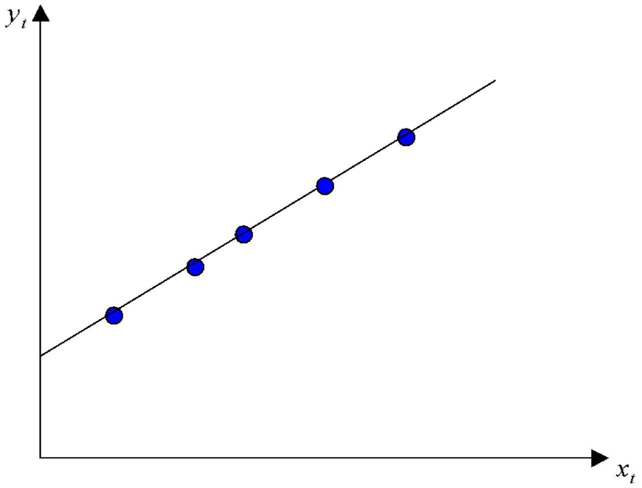
Statistical evaluation of the model

- R squared
  - That is, $TSS = ESS + RSS$
    $$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t \hat{u}_t^2$$
  - Our goodness of fit statistic is
    $$R^2 = \frac{ESS}{TSS}$$
  - But since $TSS = ESS + RSS$, we can also write
    $$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$
  - R^2 must always lie between zero and one. To understand this, consider two extremes

Statistical evaluation of the model

- R squared
  - The Limit Cases: $R^2 = 0$ and $R^2 = 1$





Statistical evaluation of the model

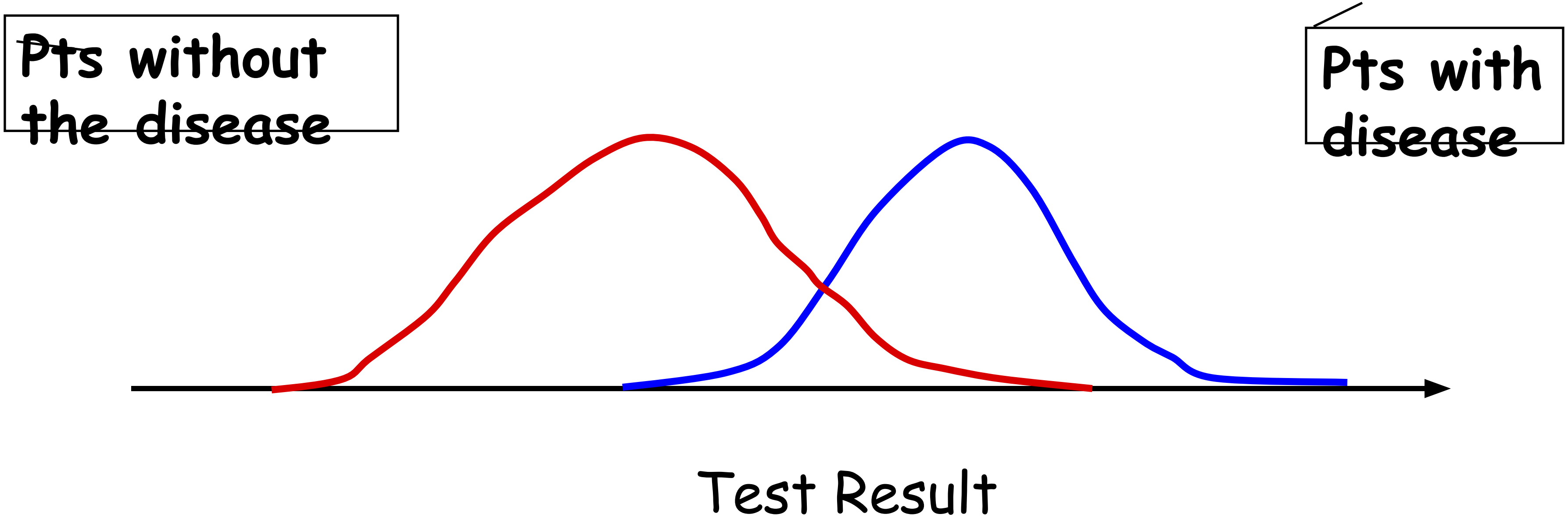
- Area under ROC curve
  - 接收者操作特徵曲線（receiver operating characteristic curve, 或者叫ROC曲線）
  - ROC曲線為一個用來呈現篩檢試驗敏感度(sensitivity)及 1-特異度(specificity)的圖形，其中 X 軸為 1-特異度，又稱為偽陽性率(false positive)，而 Y 軸為敏感度，任何一 個在曲線上的點都會對應到一個檢驗的用以區分陽性或陰性的分界點。
  - 利用曲線下的 面積(Area Under Curve, AUC)來判別 ROC 曲線的鑑別力，AUC 數值的範圍從 0 到 1，數值愈大愈好。

AUC=0.5	幾乎沒有判別力(no discrimination)
0.7≦AUC<0.8	可接受的判別力(acceptable discrimination)
0.8≦AUC<0.9	好的判別力(excellent discrimination)
AUC ≧0.9	非常好的判別力(outstanding discrimination)

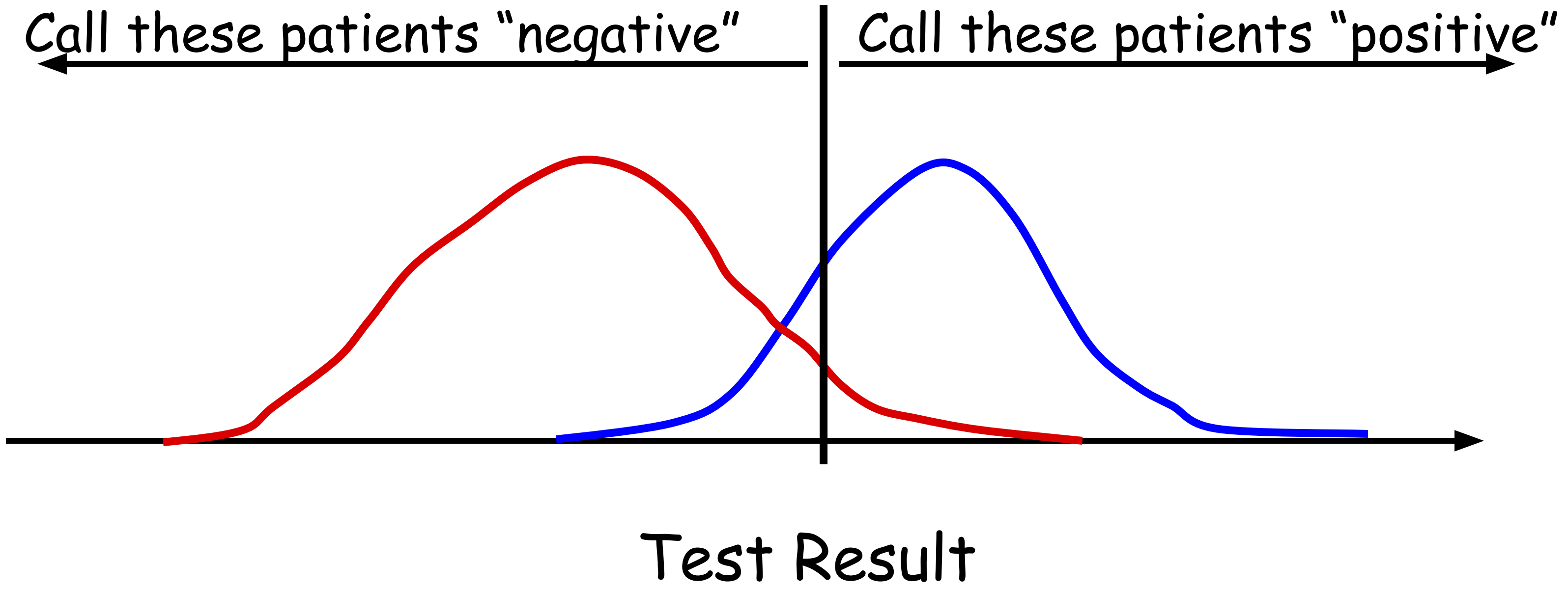
Statistical evaluation of the model

Disease \ Test	not rejected	rejected
No disease (D = 0)	☺ specificity	X Type I error (False +) α
Disease (D = 1)	X Type II error (False -) β	☺ Power 1 - β; sensitivity

Statistical evaluation of the model

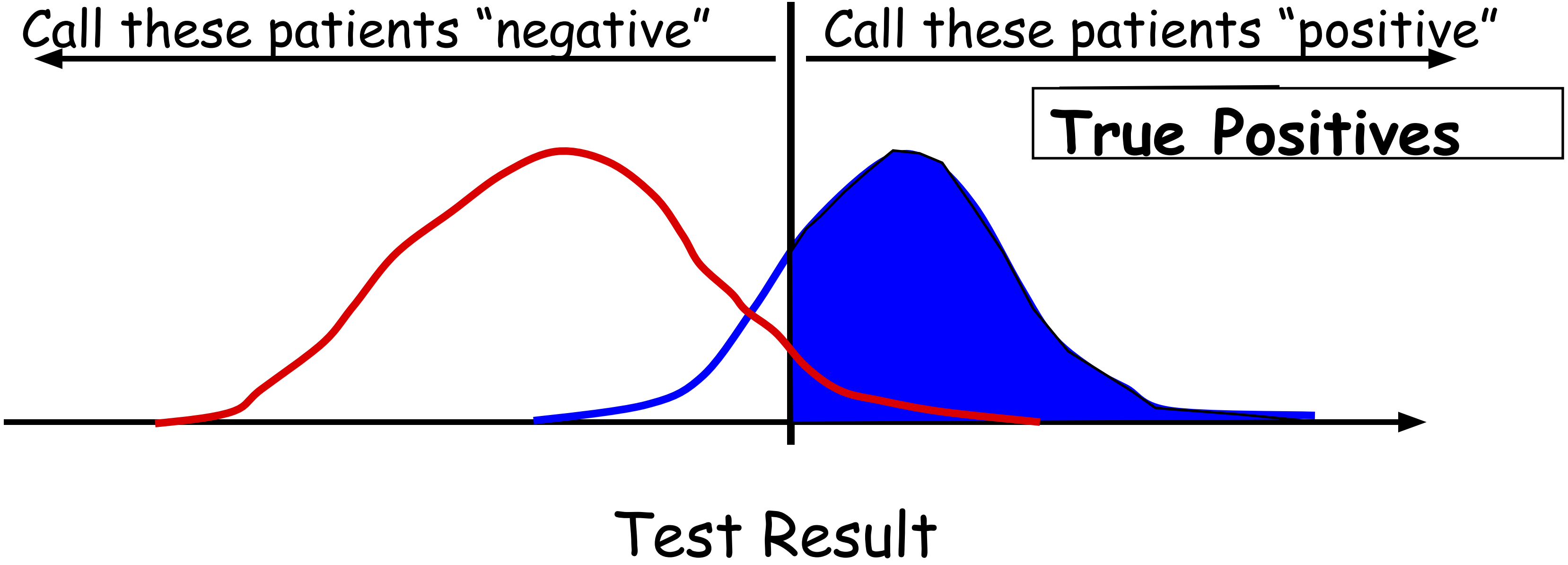


Statistical evaluation of the model



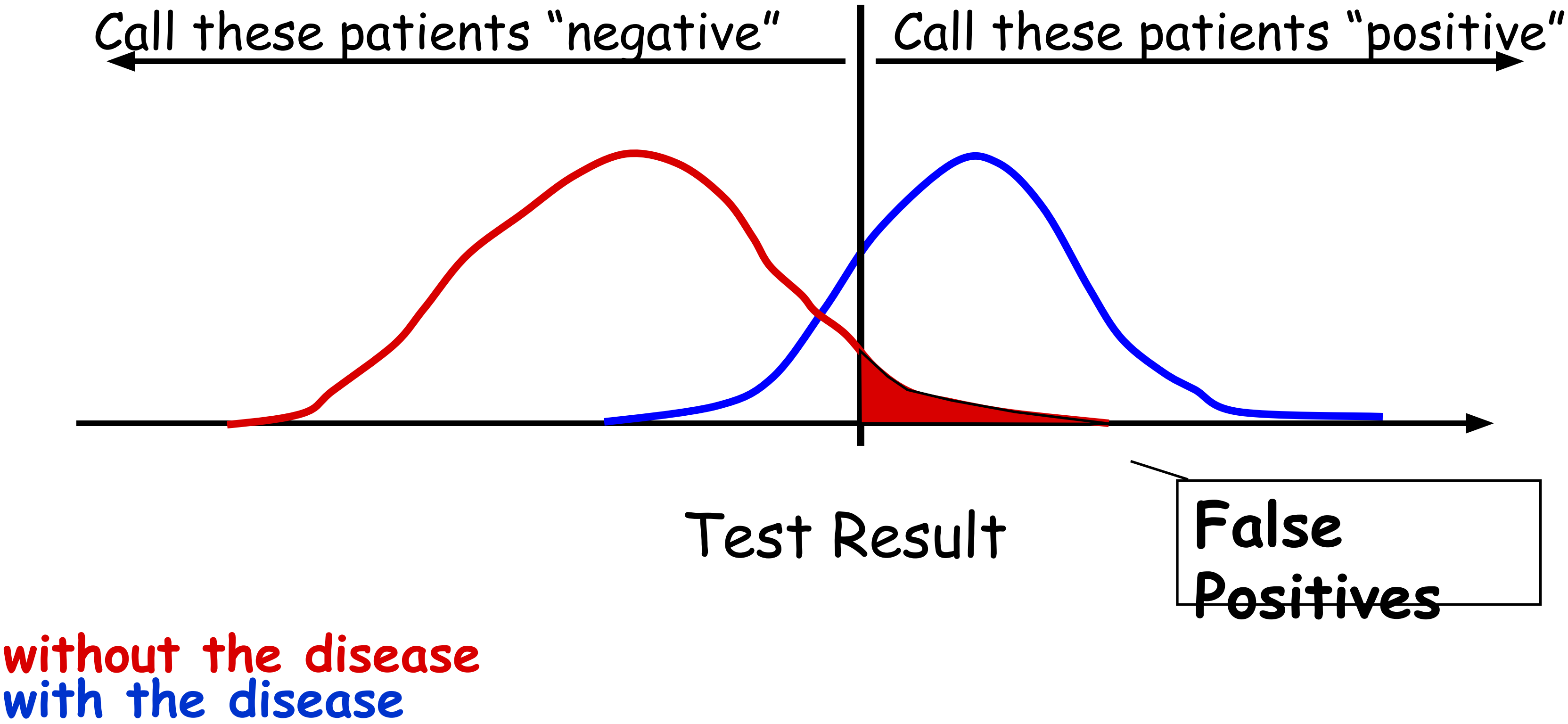
Threshold

Statistical evaluation of the model

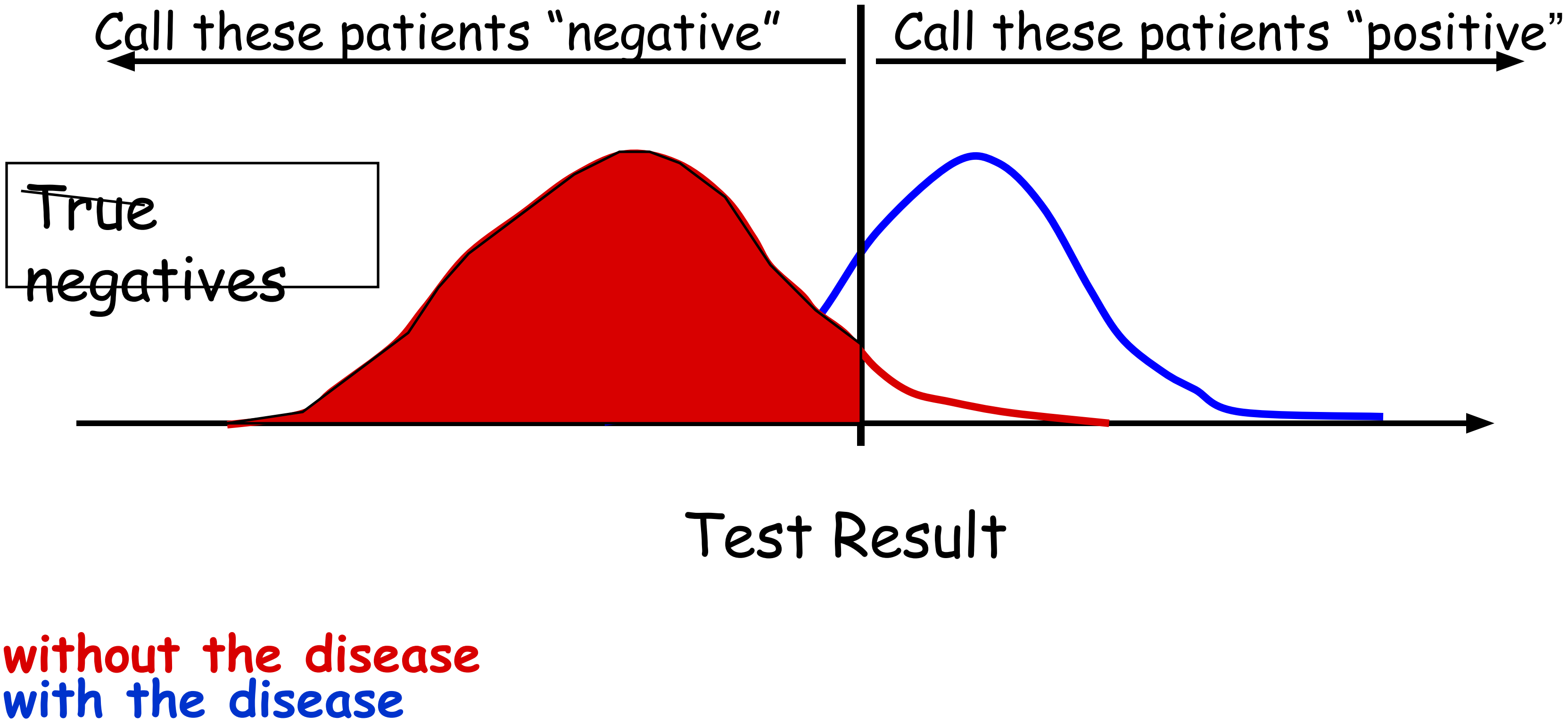


without the disease
with the disease

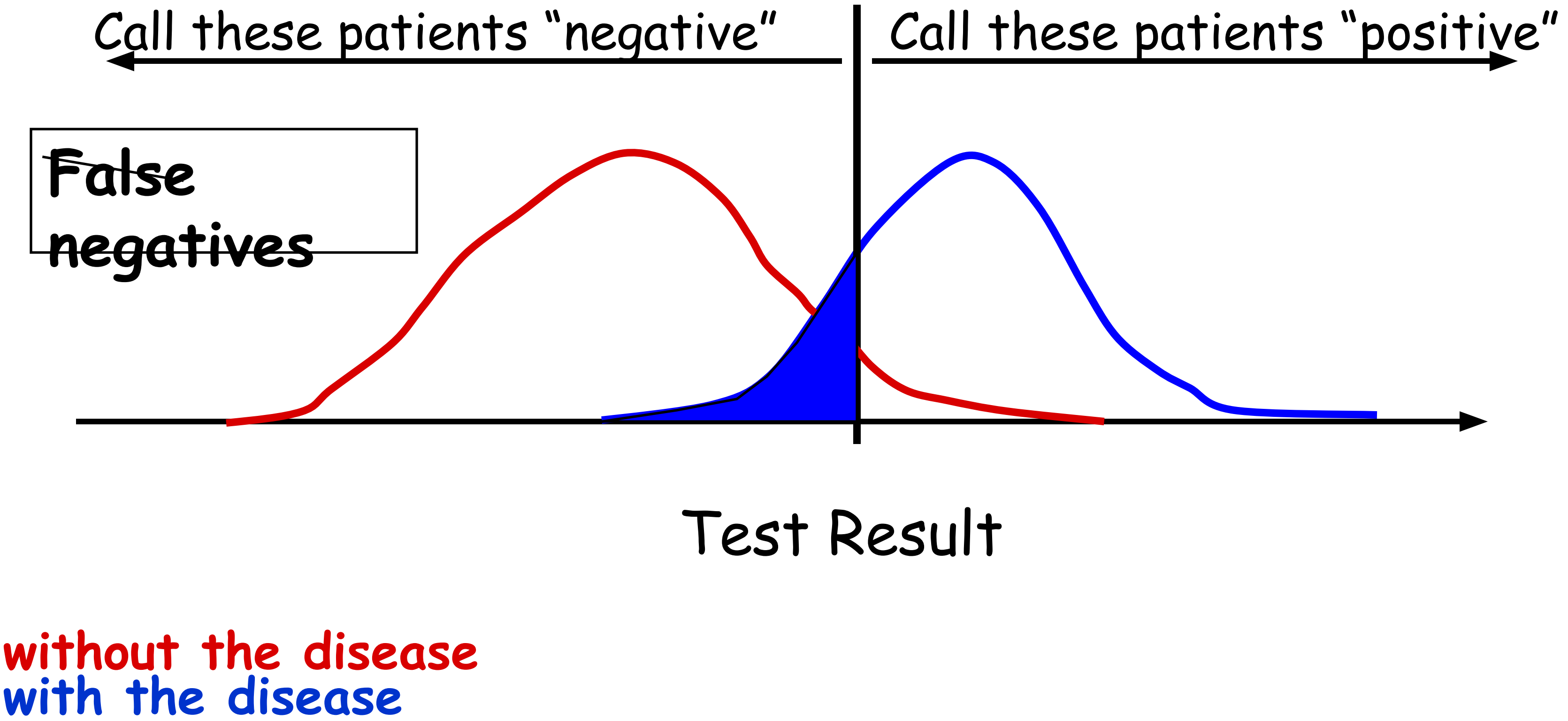
Statistical evaluation of the model



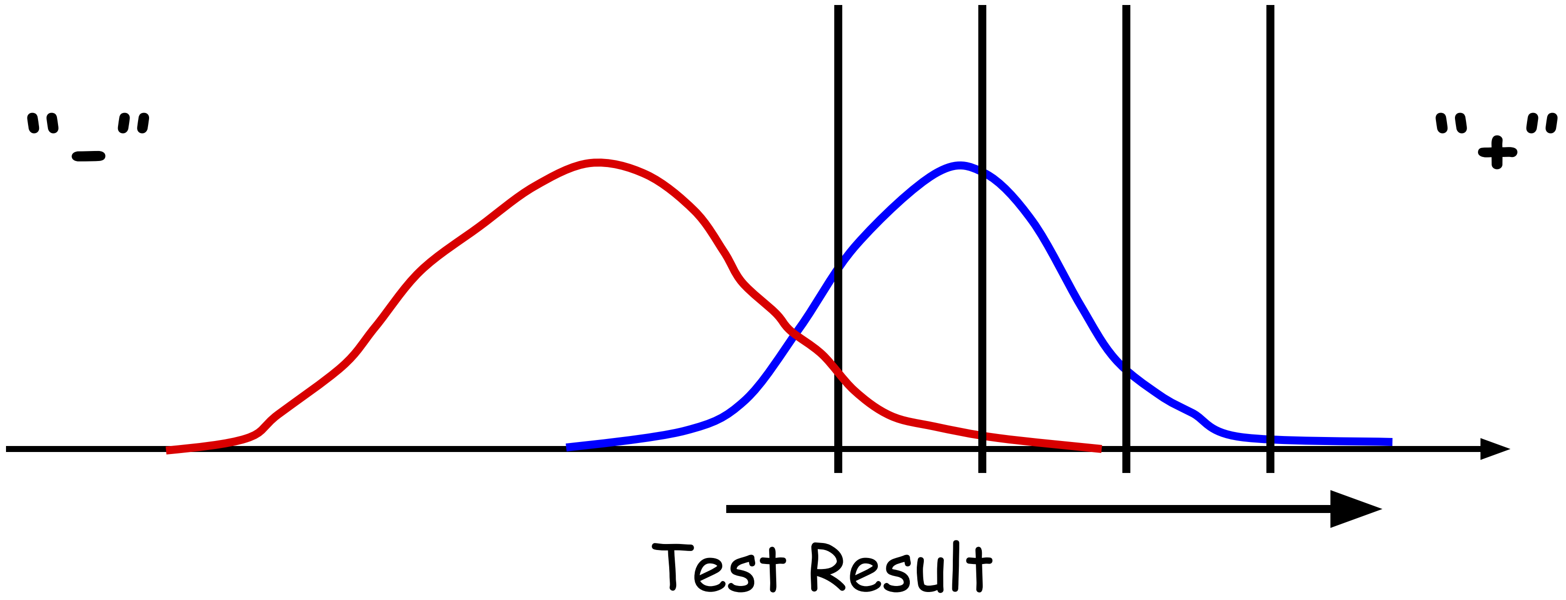
Statistical evaluation of the model



Statistical evaluation of the model



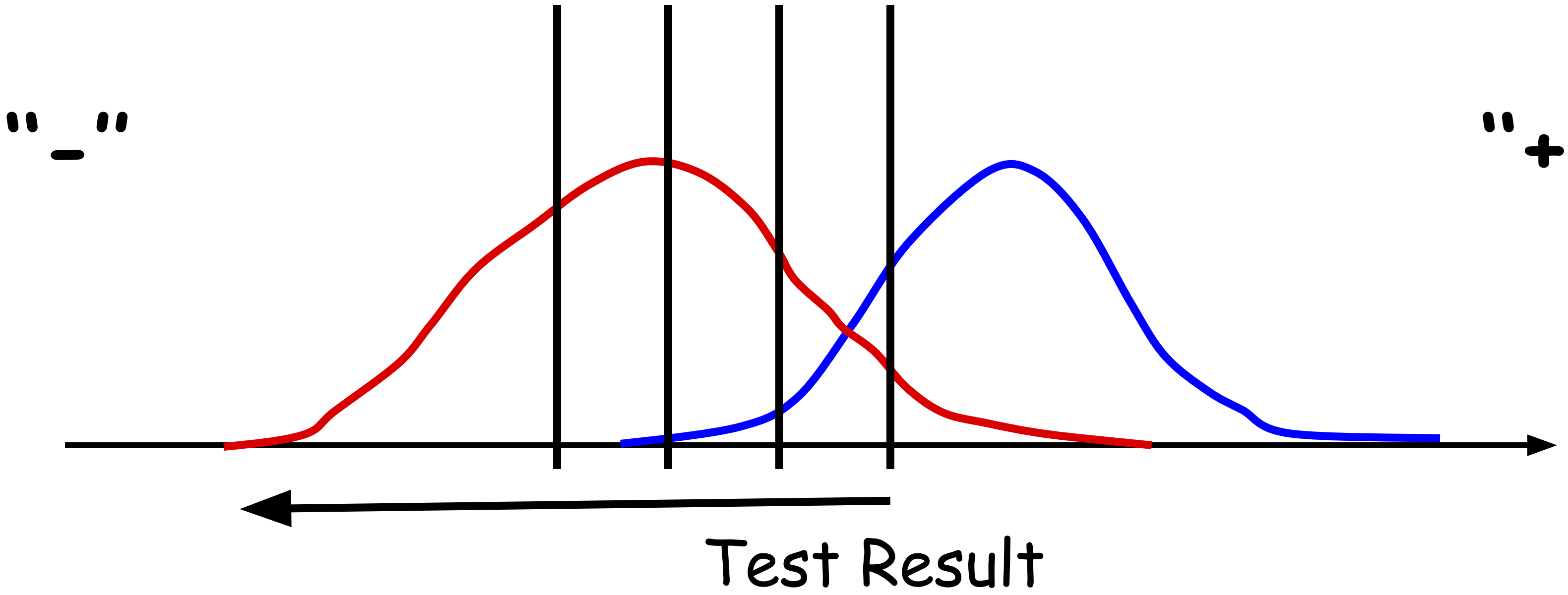
Statistical evaluation of the model



without the disease
with the disease

Moving the Threshold: right

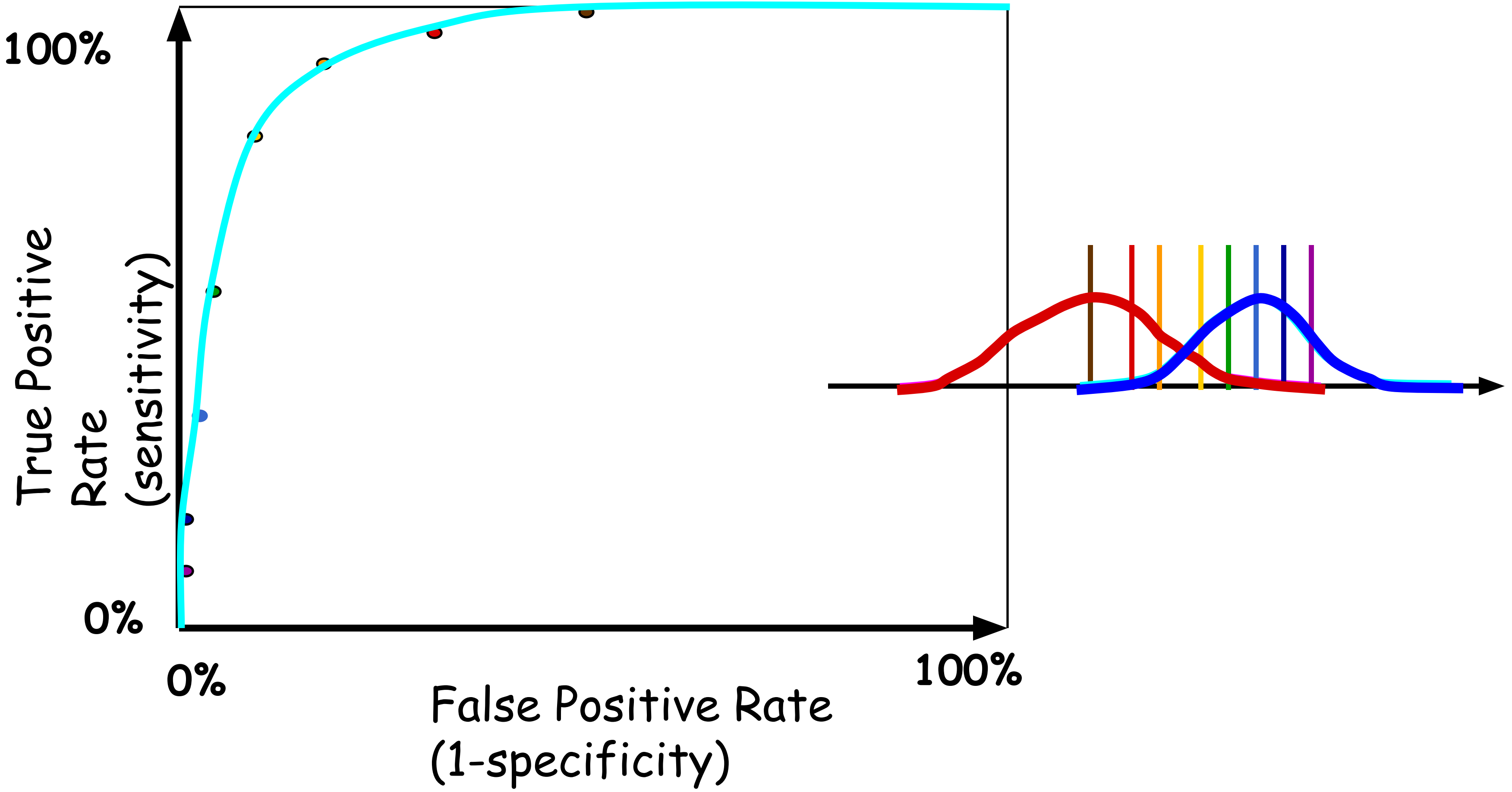
Statistical evaluation of the model



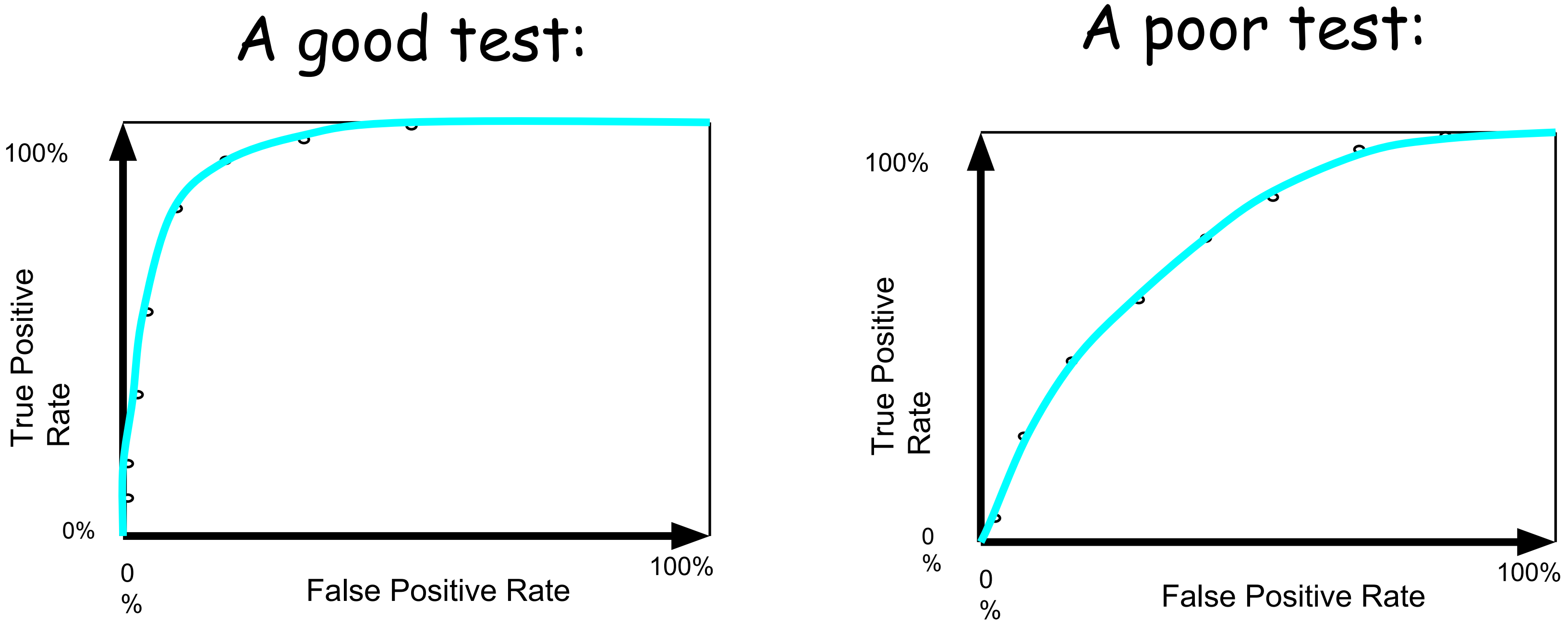
without the disease
with the disease

Moving the Threshold: left

ROC curve


100%
0%
True Positive Rate (sensitivity)
0%
100%
False Positive Rate (1-specificity)

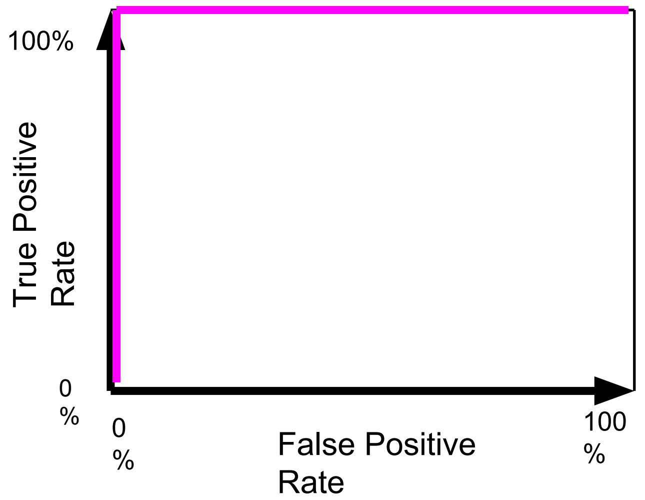
Statistical evaluation of the model



ROC curve comparison

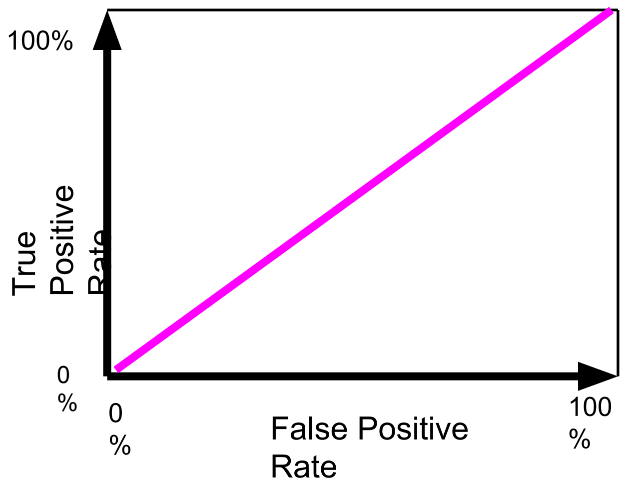
Statistical evaluation of the model

Best Test:



The distributions don't overlap at all

Worst test:



The distributions overlap completely

Statistical evaluation of the model

- Test for Pearson Correlation
    - If a linear or generalized linear model fits the data well, the predicted values and the residuals should be independent.
    - pcorr Test is to test the pearson correlation between the predicted values and the residuals.
    - Small correlation coefficient is highly desired. The 95% confidence interval of the correlation coefficient is calculated

Statistical evaluation of the model

- Pregibon Test for Linearity
    - The test is to examine the adequacy of the hypothesized link used in fitting a linear or generalized linear model.
    - http://www.inside-r.org/packages/cran/LDdiag/docs/pregibon

Statistical evaluation of the model

- Ramsey Regression Equation Specification Error Test (RESET)
  - A general specification test for the linear regression model.
  - http://www.inside-r.org/packages/cran/LDdiag/docs/ramsey

Statistical evaluation of the model

- Training Group / Validation Group
  - Randomly or chronologically assign data points to two sets d0 and d1, then train on d0 and test on d1.
  - 通常 Training Group 和 Validation Group 資料量一樣，或者可以讓 Training Group 比較大。

Example with R

```
  # Set-up:
  rm(list=ls(all=TRUE))
  ls()
  #
===========================================================================
=======
  # Data import: (if error, use 'file.choose()' instead )

  data1 <- read.csv(file="D:\\My Documents\\Studies\\101-03-18 Revise\\XY_20120318 prolong
mortality 06-09.csv", header=T) #training group: XY_20120318 prolong mortality 06-09.csv
  data2 <- read.csv(file="D:\\My Documents\\Studies\\101-03-18 Revise\\XY_20120318 prolong
mortality 10.csv", header=T) #validating group: XY_20120318 prolong mortality 10.csv

  dim(data1)
  names(data1)
```

Example with R

```
# [1] Logistic Regression Model:
sapply(data1, function(x)(sum(is.na(x))))[sapply(data1, function(x)(sum(is.na(x)))!=0]
glm.1<-glm(hospital.mortality ~
  age + gender + pre_ICU_days +section.CV + section.NS + Charlson + emergency +
readmission +Ventilator14 + Lactate14 + SOFA14 + ECMO + RRT + C1 + C2 +C3 + C4 + P1 + P2 +
P3 + P4 + N1 + N2 + N3 + N4 + G1 + G2 +M1 + M2 + E1 + E2 + E3 + R1 + R2 + O1 + O2
, family=binomial("logit")
, data=data1)
```

Example with R

```
#full model summary
summary(glm.1)
library(ROCR) # if error, type: install.packages("ROCR")
pred <- prediction( glm.1$fitted, data1$hospital.mortality)
perf <- performance(pred, "tpr", "fpr")
glm.1.auc=performance(pred, measure="auc")@y.values
glm.1.auc #area under ROC of full model (training group)

#full model validation
valid.glm.1=glm(glm.1$formula, family=binomial("logit"), data=data2)
pred <- prediction( valid.glm.1$fitted, data2$hospital.mortality)
perf <- performance(pred, "tpr", "fpr")
valid.glm.1.auc=performance(pred, measure="auc")@y.values
valid.glm.1.auc #area under ROC of full model (validating group)
```

Example with R

hospital.mortality ~ age + gender + pre_ICU_days +section.CV + section.NS + Charlson + emergency + readmission +Ventilator14 + Lactate14 + **SOFA14** + ECMO + RRT + C1 + C2 +C3 + C4 + P1 + P2 + P3 + P4 + N1 + N2 + N3 + N4 + G1 + G2 +M1 + M2 + E1 + E2 + E3 + R1 + R2 + O1 + O2

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5587  -0.6795  -0.3952   0.5679   2.5124

Coefficients:
               Estimate Std. Error z value     Pr(>|z|)
(Intercept)   -4.594893   0.488068  -9.414      < 2e-16 ***
age            0.003505   0.005062   0.692     0.488696
gender         0.342336   0.157070   2.180     0.029293 *
pre_ICU_days   0.005480   0.002518   2.176     0.029518 *
section.CV    -0.826303   0.232870  -3.548     0.000388 ***
section.NS    -0.682307   0.273851  -2.492     0.012720 *
Charlson       0.098791   0.031971   3.090     0.002001 **
emergency     -0.687626   0.161074  -4.269  0.000019634 ***
readmission    0.043324   0.196597   0.220     0.825582
Ventilator14   0.462103   0.232329   1.989     0.046701 *
Lactate14      0.535595   0.100972   5.304  0.000000113 ***
SOFA14         0.317750   0.029293  10.847      < 2e-16 ***
ECMO          -0.310195   0.311449  -0.996     0.319263
RRT           -0.074582   0.199998  -0.373     0.709210
C1           -12.138925 614.619337  -0.020     0.984243
C2             1.067361   0.440488   2.423     0.015387 *
C3             0.902967   0.370864   2.435     0.014901 *
C4            -0.061527   0.485290  -0.127     0.899112
P1             0.578752   0.213117   2.716     0.006615 **
P2             0.711588   0.407194   1.748     0.080544 .
P3             0.466800   0.200678   2.326     0.020012 *
P4            -0.052740   0.312925  -0.169     0.866160
N1             0.826024   0.313789   2.632     0.008478 **
N2             1.286522   0.340749   3.776     0.000160 ***
N3             1.821515   1.264866   1.440     0.149843
N4            -0.962361   2.014840  -0.478     0.632909
G1            -0.310196   0.452273  -0.686     0.492801
G2             0.775057   0.346873   2.234     0.025456 *
M1             0.464924   0.373047   1.246     0.212659
M2             0.018320   0.793575   0.023     0.981582
E1             0.689336   0.189340   3.641     0.000272 ***
E2             0.134021   0.438578   0.306     0.759923
E3             1.426109   1.302409   1.095     0.273526
R1            -0.128089   0.373997  -0.342     0.731984
R2           -12.211639 882.743433  -0.014     0.988963
O1           -12.461044 455.961250  -0.027     0.978197
O2             1.005848   0.534153   1.883     0.059691 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1734.0  on 1345  degrees of freedom
Residual deviance: 1170.8  on 1309  degrees of freedom
AIC: 1244.8
```

area under ROC of full model (training group) = 0.8565124

area under ROC of full model (validating group) = 0.906011

Example with R

1. Backward logistic regression (Backward AIC)

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.9533608  0.8292012   2.356 0.018487 *
gender       0.4460494  0.1617067   2.758 0.005809 **
pre_ICU_days 0.0064838  0.0023701   2.736 0.006224 **
section.CV  -1.0210690  0.2385238  -4.281 1.86e-05 ***
section.NS  -1.2630962  0.3077589  -4.104 4.06e-05 ***
Charlson     0.1100117  0.0322210   3.414 0.000639 ***
emergency   -0.6700438  0.1618625  -4.140 3.48e-05 ***
GCS14       -0.1798783  0.0296179  -6.073 1.25e-09 ***
MAP14       -0.0271799  0.0058700  -4.630 3.65e-06 ***
IE14         0.0393042  0.0162465   2.419 0.015553 *
Ventilator14 0.3666970  0.2332128   1.572 0.115865
Lactate14    0.5509356  0.1046810   5.263 1.42e-07 ***
Platate14   -0.0018999  0.0007068  -2.688 0.007188 **
SOFA14       0.1462924  0.0360471   4.058 4.94e-05 ***
RRT          0.4814565  0.2065635   2.331 0.019764 *
C2           0.7345285  0.3988000   1.842 0.065498 .
C3           0.6769529  0.3743786   1.808 0.070575 .
P1           0.3728354  0.1878151   1.985 0.047131 *
N1           0.4622248  0.3128482   1.477 0.139549
N2           0.9067436  0.3515911   2.579 0.009909 **
G2           0.7441743  0.3414040   2.180 0.029276 *
E1           0.6377514  0.1801624   3.540 0.000400 ***
O2           0.8285994  0.5320020   1.558 0.119349

AIC: 1163.2
area under ROC of backward AIC model (training group) = 0.8716353
area under ROC of backward AIC model (validating group) = 0.9092957
```

Example with R

2. LASSO (or Elastic Net) model*

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.2463096  0.8377257   2.681 0.007331 **
gender       0.4281883  0.1613063   2.655 0.007943 **
pre_ICU_days 0.0059957  0.0023510   2.550 0.010763 *
section.CV  -0.8727713  0.2256852  -3.867 0.000110 ***
section.NS  -1.0614538  0.2677569  -3.964 7.36e-05 ***
Charlson     0.1147350  0.0322277   3.560 0.000371 ***
emergency   -0.6391903  0.1610816  -3.968 7.24e-05 ***
GCS14       -0.1930284  0.0300467  -6.424 1.32e-10 ***
MAP14       -0.0266157  0.0058492  -4.550 5.36e-06 ***
IE14         0.0481174  0.0167587   2.871 0.004089 **
Ventilator14 0.3895401  0.2341235   1.664 0.096148 .
Lactate14    0.5324212  0.1049901   5.071 3.95e-07 ***
Bilirubin14  0.0222889  0.0164170   1.358 0.174568
Platate14   -0.0018687  0.0007029  -2.659 0.007843 **
SOFA14       0.1217272  0.0401024   3.035 0.002402 **
RRT          0.4935298  0.2133360   2.313 0.020701 *
C3           0.4681163  0.3586186   1.305 0.191780
P1           0.2865769  0.1826374   1.569 0.116624
N2           0.7188470  0.3363666   2.137 0.032590 *
G2           0.5878637  0.3357123   1.751 0.079930 .
M1           0.4069994  0.3861730   1.054 0.291915
E1           0.5111870  0.1718532   2.975 0.002934 **

AIC: 1165
area under ROC of LASSO model (training group) = 0.8702767
area under ROC of LASSO model (validating group) = 0.8993478
```

Example with R

3. Ridge regression model with a stepwise variable selection

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.948343   0.535632  -1.771   0.0766 .
SOFA14             0.341082   0.026999  12.633  < 2e-16 ***
Lactate14          0.546427   0.094608   5.776 7.66e-09 ***
MAP14             -0.026524   0.005323  -4.982 6.28e-07 ***
Charlson           0.139717   0.029482   4.739 2.15e-06 ***
SOFA14:emergency  -0.126853   0.021810  -5.816 6.02e-09 ***

AIC: 1236.4
area under ROC of step-ridge regression model (training group) = 0.842984
area under ROC of step-ridge regression model (validating group) = 0.8508282
```