R：Statistical Programming Methods
R：程式、機率與統計

Logistic Regression

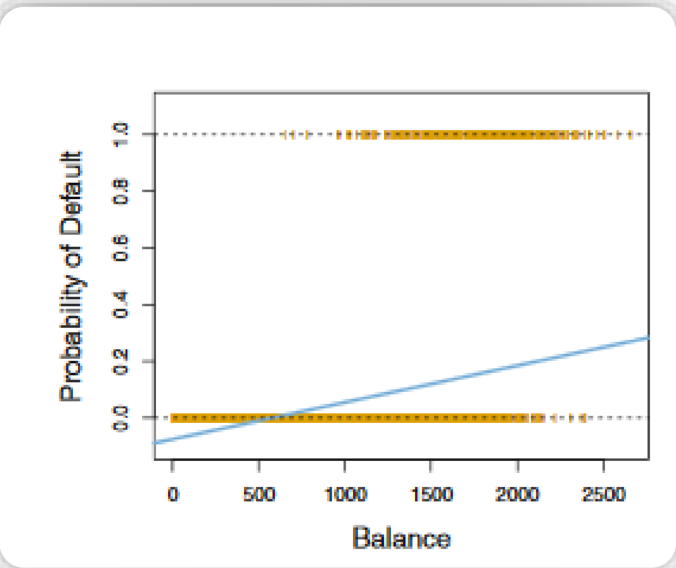
Introduction

- Previous section uses linear regression to examine and predict the outcome of **continuous variables**.

- What if data is binary? (e.g., Yes/No; A/B)
- What we are interested in? (what do we want to predict?)
- p: the probability of getting 1 given some variable(s)

Can we use linear regression?

- NO.
- If we used $p = \beta_0 + \beta_1 X_1$, the probability p may be negative or higher than one (both are impossible!)
- The probabilities need to fall between **0 and 1.**



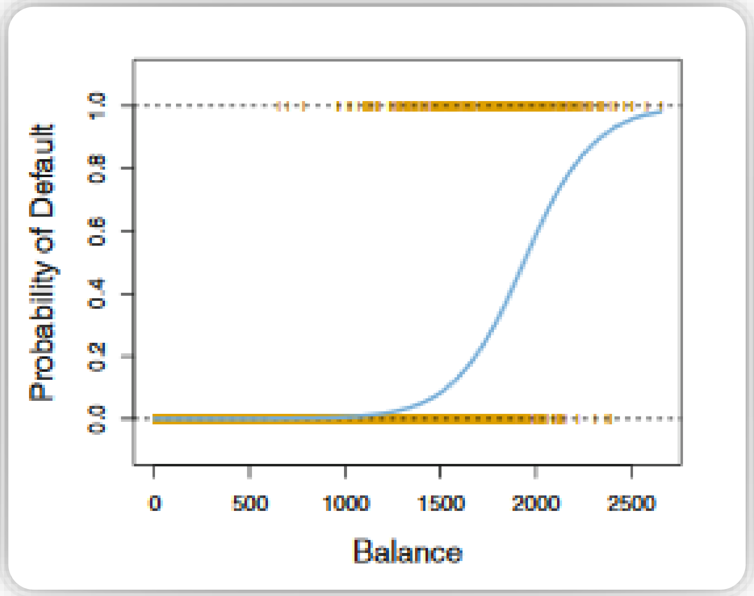
Logit Function

- Converting using logistic response / inverse logit function
- $p = \frac{1}{1+ e^{-(\beta_0+\beta_1 \mathrm{X}_1)}}$ (ensuring that p stays between 0 and 1)
- $Odds(y = 1) = \frac{p}{1-p}$ (and $p = \frac{Odds}{1+Odds}$)
- $Odds(y = 1) = e^{(\beta_0+\beta_1 \mathrm{X}_1)}$
- $\log(Odds(y = 1)) = \beta_0 + \beta_1 \mathrm{X}_1$

Logistic Function

- We need a function that outputs a number between 0 and 1.
- We use the logistic function instead:

$$P(Y) = \frac{e^{\beta 0+\beta 1X1}}{1+\ e^{\beta_0+\beta_1 X_1}}$$



Example – Heart Disease Data

- sex: male or female(Nominal, male=1, female=0)
- Age: Age of the patient
- Current Smoker: whether or not the patient is a current smoker (Nominal)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day
- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)
- Tot Chol: total cholesterol level (Continuous)
- Sys BP: systolic blood pressure (Continuous)
- Dia BP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)
- Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.
- Glucose: glucose level (Continuous)
- 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

```r
glm_heart <- glm(TenYearCHD~male+age+currentSmoker+
                 cigsPerDay+totChol+diaBP+BMI+heartRate, data=df, family="binomial")
summary(glm_heart)
```

$$\log(Odds(y=1)) = \beta_0 + \beta_1 X_1$$

```
Call:
glm(formula = TenYearCHD ~ male + age + currentSmoker + cigsPerDay +
    totChol + diaBP + BMI + heartRate, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5748  -0.6134  -0.4393  -0.2916   2.8302

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.236511   0.573108 -16.117  < 2e-16 ***
male           0.457013   0.099676   4.585 4.54e-06 ***
age            0.079565   0.005865  13.565  < 2e-16 ***
currentSmoker  0.049558   0.146075   0.339 0.734410
cigsPerDay     0.020175   0.005769   3.497 0.000471 ***
totChol        0.002391   0.001032   2.318 0.020470 *
diaBP          0.022659   0.003948   5.739 9.52e-09 ***
BMI            0.011932   0.011720   1.018 0.308621
heartRate      0.002085   0.003865   0.539 0.589612
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3503.5  on 4139  degrees of freedom
Residual deviance: 3161.7  on 4131  degrees of freedom
  (因為不存在，98 個觀察量被刪除了)
AIC: 3179.7

Number of Fisher Scoring iterations: 5
```

- $e^{0.079} = 1.082$,
- For every year older, the odds of getting heart disease is 1.082 higher.
- The probability of getting heart disease than NOT getting heart disease

How good is the model?

- Independence of variables and multicollinearity assumptions still apply
- Use deviance and AIC to measure the goodness-of-fit (Smaller is better!)
  - Deviance (模型偏差)
  - AIC (Akaike's Information Criterion)
- Use prediction to see if the model is good or not

Prediction – Confusion Matrix

```
#Separate the data into 80% training and 20% testing
train = sample(1:nrow(df), nrow(df)*0.8)
training_df = df[train,]
testing_df = df[-train,]

train_heart <- glm(TenYearCHD~male+age+currentSmoker+

cigsPerDay+totChol+diaBP+BMI+heartRate, data=training_df,
family="binomial")
```

Prediction – Confusion Matrix

```
predict_heart <- predict(train_heart, newdata=testing_df,
type="response")
library(regclass)
 confusion_matrix(train_heart, testing_df)
 ##              Predicted 0 Predicted 1 Total
 ## Actual 0             680           3   683
 ## Actual 1             134           6   140
 ## Total                814           9   823
```

	Predicted 0	Predicted 1
Actual 0	680	3
Actual 1	134	6

Accuracy: $\frac{680+6}{680+3+134+6} = 82.25\%$,
Misclassification: $1 - 0.8225 = 17.75\%$

Prediction – Confusion Matrix

- Precision (精準度)
  - the accuracy of a predicted positive outcome
  - 預測為1，也真的為1
  - $\frac{\mathrm{TP}}{\mathrm{TP+FP}} = \frac{6}{3+6} = 67\%$
  - Type 1 errors
- Recall (召回率)
  - the strength of the model to predict a positive outcome
  - 實際為1，預測也為1
  - $\frac{\mathrm{TP}}{\mathrm{TP+FN}} = \frac{6}{6+134} = 4.29\%$
  - Type 2 error

	Predicted 0	**Predicted 1**
Actual 0	680 (TN)	3 (FP)
Actual 1	134 (FN)	6 (TP)

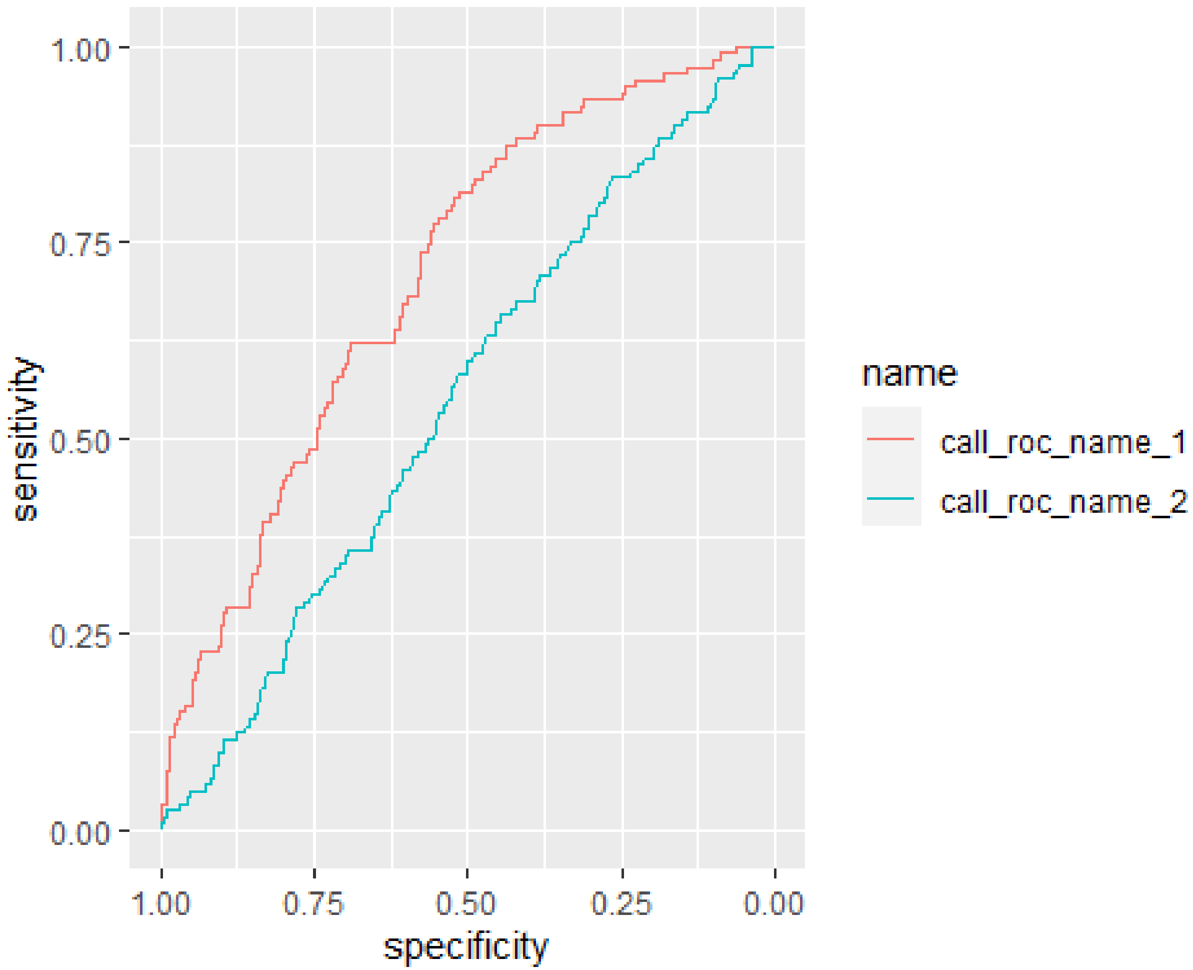
F1 Score

- As always increasing type I errors will decrease type II and decreasing type I will increase type II.
- Harmonic mean between precision and recall
- $\frac{2(TP)}{2(TP)+FP+FN} = \frac{2(6)}{2(6)+3+134} = 0.08$
- Numbers closer to one show good precision AND recall

```r
train2_heart <- glm(TenYearCHD~BMI, data=training_df,
family="binomial")
predict2_heart <- predict(train2_heart, newdata=testing_df,
type="response")

testing_df$predict1 <- predict_heart
testing_df$predict2 <- predict2_heart
```

```r
library(pROC)
rocobj1 <- roc(testing_df$TenYearCHD, testing_df$predict1)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
rocobj2 <- roc(testing_df$TenYearCHD, testing_df$predict2)
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
ggroc(list(call_roc_name_1 = rocobj1, call_roc_name_2 = 
rocobj2))
```


1.00
0.75
0.50
0.25
0.00
sensitivity
1.00
0.75
0.50
0.25
0.00
specificity
name
call_roc_name_1
call_roc_name_2