# 4. Prediction

## Prediction

In the `glmdr` framework, we compute the weighted estimated probability $\hat{p^*}$ for the prediction. We generate two datasets which combine of training sets and $i$-th observation of testing sets with $y = 0$ and $y = 1$ separately. We fit either `glm` or `glmdr` model depending on the presence of the complete separation then calculate weighted estimated probability using the model (Akaike) weights, $\hat{p^*} = w_0 * \hat{p_0} + w_1 * \hat{p_1}$ where $w_i = \exp(-\frac{IC_i}{2})/(\exp(-\frac{IC_1}{2}) + \exp(-\frac{IC_2}{2}))$ and $IC$ is the information criterion such as AIC, AICc and BIC.

## Endometrial

### 1. Generate datasets

In this experiment, we use 70% of data as a training set and 30% of data as a testing set. We are going to construct two datasets based on the training set. Each data set has additional one data point from testing set with $y = 0$ and $y = 1$ separately.

```r
training_set_size <- 0.7
n <- nrow(endometrial)
idx <- c(rep(0,n*(1-training_set_size)),rep(1,n*training_set_size))
set.seed(528)
idx <- sample(idx)
training <- endometrial[idx==1,]
testing <- endometrial[idx!=1,]
testing_X <- endometrial[idx!=1,-4]
testing_Y <- endometrial[idx!=1,4]
i <- 3
new_0 <- cbind("HG"=0,testing_X[i,])
new_1 <- cbind("HG"=1,testing_X[i,])
new_0 <- rbind(training,new_0)
new_1 <- rbind(training,new_1)
cbind(tail(new_0,5),tail(new_1,5))
```

```
##    NV PI   EH HG NV PI   EH HG
## 76  1 21 0.98  1  1 21 0.98  1
## 77  0  5 0.35  1  0  5 0.35  1
## 78  1 19 1.02  1  1 19 1.02  1
## 79  0 33 0.85  1  0 33 0.85  1
## 5   0 20 1.28  0  0 20 1.28  1
```

We can see two datasets consist of training set and new data point with different values in the reponse variable.

### 2. Modeling and Estimating $\hat{p}$

We fit the `glmdr` model and check if 1) the complete separation presents and 2) the linearity of the new data point (if the complete separation exists). After all, there will be 3 cases:

1. Complete separation does not exist, then we can use original model.

2. Complete separation exists and new data point is the problematic point (`linearity == TRUE`), then use inference() function.
3. Complete separation exists and new data point is not the problematic point, then use original model.

### Case 1

We can see complete separation presents in the both models but new data points are not problematic points. Thus, we can use the `glm` function to compute the $\hat{p}$.

```r
new_0_mod <- glmdr(HG~.,data=new_0,family="binomial")
tail(!new_0_mod$linearity,1)
```

```
##     5
## FALSE
```

```r
phat_0 <- tail(predict(new_0_mod$om,type="response"),1)
```

```r
new_1_mod <- glmdr(HG~.,data=new_1,family="binomial")
tail(!new_1_mod$linearity,1)
```

```
##     5
## FALSE
```

```r
phat_1 <- tail(predict(new_1_mod$om,type="response"),1)
cbind(phat_0,phat_1)
```

```
##      phat_0    phat_1
## 5 0.3950498 0.4369958
```

### Model weights

Now, we calculate the weighted of two models to obtain weighted estimated probability, $\hat{p^*}$. Model weight is defined as:

$$w_i = \frac{\exp(-\frac{IC_i}{2})}{\exp(-\frac{IC_1}{2}) + \exp(-\frac{IC_2}{2})}$$

where $IC$ represents the information criterion (* sum of two weights must be 1).

```r
crit1 <- exp(-AIC(new_0_mod$om)/2)
crit2 <- exp(-AIC(new_1_mod$om)/2)
w0 <- crit1/(crit1+crit2)
w1 <- crit2/(crit1+crit2)
w0+w1
```

```
## [1] 1
```

### Weighted estimated probability

Based on the model weighted we obtained in the previous part, we compute the weighted estimated probability:

$$\hat{p^*} = w_0 * \hat{p_0} + w_1 * \hat{p_1}$$

.

```
phat_star <- w0 * phat_0 + w1 * phat_1
phat_star
```

```
##         5
## 0.4124921
```

### Construct Confidence Interval

We use Wilson (1927) confidence interval.

$$CI_w = (\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z^2_{\alpha/2} * \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2_{\alpha/2}}{4n^2}})/(1 + \frac{z^2_{\alpha/2}}{n})$$

.

```
nn <- nrow(new_0)
z_score <- qnorm(1-0.05/2)

Wilson_upper <- (phat_star + z_score^2/(2*nn) +  z_score * sqrt((phat_star*(1-phat_star)/nn)+(z_score^2/
Wilson_lower <- (phat_star + z_score^2/(2*nn) -  z_score * sqrt((phat_star*(1-phat_star)/nn)+(z_score^2/
cbind(Wilson_lower,Wilson_upper)
```

```
##   Wilson_lower Wilson_upper
## 5    0.2941958    0.5418386
```

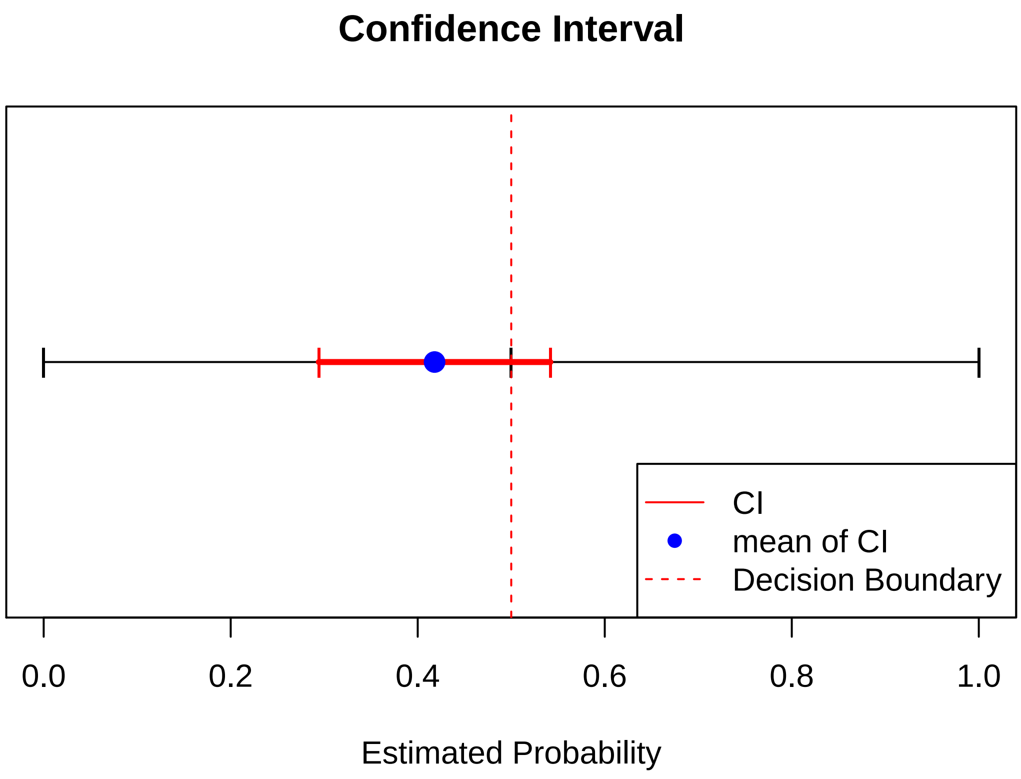
```
binom.confint(x=phat_star*nn,n=nn,methods="wilson")
```

```
##   method        x  n      mean     lower     upper
## 1 wilson 23.51205 57 0.4124921 0.2941958 0.5418386
```

### Prediction

$$\hat{y} = \begin{cases} 1\,, & \frac{CI_{upper}+CI_{lower}}{2} >= 0.5 \\ 0, & \frac{CI_{upper}+CI_{lower}}{2} < 0.5 \end{cases}$$

```
x <- c(0,0.5,1)
x <- data.frame(x,0.1)
x1 <- c(Wilson_lower,Wilson_upper)
x1 <- data.frame(x1,0.10)
plot(x, type = 'o', pch = '|', ylab = '',yaxt='n', main="Confidence Interval",xlab="Estimated Probabili
lines(x1, type = 'o', pch = '|', ylab = '',col="red",lwd=3)
abline(v=0.5,col="red",lty=2)
points(mean(c(Wilson_lower,Wilson_upper)),0.1,col="blue",pch=16,cex=1.5)
legend("bottomright", legend=c("CI", "mean of CI", "Decision Boundary"),
       col=c("red", "blue","red"), lty=c(1,NA,2), pch=c(NA,16,NA),cex=1)
```



**Reference**

*Dasgupta, A., Cai, T. T., & Brown, L. D. (2001). Interval Estimation for a Binomial Proportion. Statistical Science, 16(2), 101-133. doi:10.1214/ss/1009213286*

*Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference - 2nd ed.: A practical information-theoretic approach. New york, ny: Springer-verlag new york.*

*Dorai-Raj S (2009) Binom: binomial confidence intervals for several parameterizations. http://carn.r-project.org/package=binom.*