GLM for Bernoulli and Binomial Response

1. Statistical model

$y_i \sim Bernoulli(\pi_i)$

$y_i \sim Binomial(n_i, \pi_i)$

2. Link function

$g(\pi_i)$: use $b(\theta)$ in the standard Exponential family form.

a. Logit:

$$log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}}{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}} + 1}$$

b. Probit:

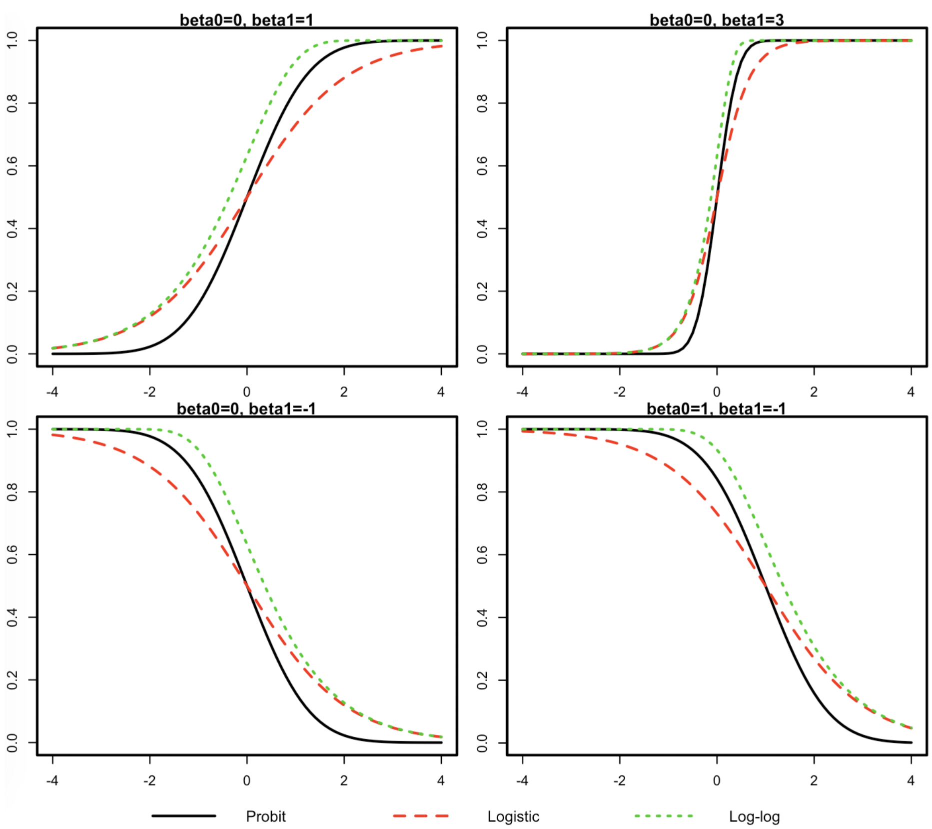
$$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\pi_i = P\{N(0,1) \leq \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}\}$$

c. Complementary-log-log (cloglog):

$$log(-log(1-\pi)) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

$$\pi_i = 1 - exp\{-exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})\}$$



3. Estimation & Inference

a. Maximum likelihood estimation $\rightarrow \hat{\beta}, \hat{\pi}$

i. Likelihood

$$L = \prod_{i=1}^{n} \binom{n_i}{Y_i} \cdot (\frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}}{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}+1})^{Y_i}(\frac{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}}{e^{\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}}+1})^{n-Y_i}$$

Where: $\begin{cases} n_i, Y_i, X_i & -observed \\ \beta_0, \beta_1, \beta_1 & -parameter \end{cases}$

ii. Find $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ to maximize L

iii. Estimate $\hat{\pi}$

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{i2}}}{e^{\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\hat{\beta}_2 X_{i2}}+1}$$

$\because \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are MLE

$\therefore \hat{\pi}$ is also a MLE

b. Confidence Interval

i. $(1-\alpha)\%$ CI for β_1:

$$\hat{\beta}_1 \pm Z_{1-\frac{\alpha}{2}} \cdot se(\hat{\beta}_1)$$

$\triangle$ use $1-\frac{\alpha}{g}$ as the confidence level for each interval when there are g CIs in the family.

ii. $(1-\alpha)\%$ CI for e^{β_1}:

(1) Find CI for β_1:

$$(\underbrace{\hat{\beta}_1 - Z_{1-\frac{\alpha}{2}} \cdot se(\hat{\beta}_1)}_{L\beta_1}, \underbrace{\hat{\beta}_1 + Z_{1-\frac{\alpha}{2}} \cdot se(\hat{\beta}_1)}_{U\beta_1})$$

(2) CI for β_1:

$$(e^{L\beta_1}, e^{U\beta_1})$$

(3) If X_1 increases k units, CI for odds-ratio(e^{β_1}):

$$(e^{kL\beta_1}, e^{kU\beta_1}) \text{ , or}$$
$$((e^{L\beta_1})^k, (e^{U\beta_1})^k)$$

c. Test hypothesis of one β

$$log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- $\begin{cases} H_0 : \beta_1 = b_1 \\ H_A : \beta_1 \neq b_1 \end{cases}$

- Test statistics:
$$Z_{obs} = \frac{\hat{\beta}_0 - b_1}{se(\hat{\beta}_1)}$$

- p-value $= 2 \cdot P(Z > |Z_{obs}|)$
- If $\beta_1 = 0$ ($e^{\beta_1} = 1$), then X_1 and π are not associated (X_1 and log-odds are not linearly associated)

d. Test hypothesis of several βs

$$log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (Full)$$
$$log\frac{\pi}{1-\pi} = \beta_0 \quad (Reduced)$$

- $\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_A : ALOI \end{cases}$

- Likelihood Ratio Test (Deviance test):

$$\Lambda = \frac{\max\limits_{H_0} Likelihood}{\max\limits_{H_0 \bigcup H_A} Likelihood} = \frac{Likelihood(Reduced)}{Likelihood(Full)}$$
$$G^2 = -2log\Lambda = -2log\frac{Likelihood(Reduced)}{Likelihood(Full)}$$
$$= \underbrace{-2log(Likelihood(Reduced))}_{deviance(R)} - \underbrace{(-2log(Likelihood(Full)))}_{deviance(F)}$$
$$G^2 \underset{n\to\infty}{\overset{H_0}{\sim}} \chi^2_{(df=p_2-p_1)}$$
$$\begin{cases} p_1 = \#\text{ of parameters in reduced model,} \\ p_2 = \#\text{ of parameters in full model} \end{cases}$$

In GLM, Deviance $= -2log(Likelihood(\hat{\beta}_{MLE}))$

- p-value $= P(\chi^2_{df} > G^2)$

4. Interpretation

True: $log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Interpretation for β_1: After adjusting for the effect of the other predictors (X_2), as X_1 increases **k** unit, then

a. the **log-odds** will increase **k**β_1.
When $X_1 = a$,

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = log\frac{\pi_0}{1-\pi_0} = \text{log-odds}_0$

When $X_1 = a + \mathbf{k}$,

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = log\frac{\pi_0}{1-\pi_0} + \mathbf{k}\beta_1 = \text{log-odds}_1$

$\rightarrow$ Change: $\text{log-odds}_1 - \text{log-odds}_0 = \mathbf{k}\beta_1$

b. the **log-odds-ratio** will increase $\mathbf{k}\beta_1$.

When $X_1 = a$,

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = \text{log-odds}_0 = log(odds_0)$

$\text{log-odds-ratio}_0 = \frac{\text{log-odds}_0}{\text{log-odds}_0} = log(odds_0) - log(odds_0) = 0$

When $X_1 = a + \mathbf{k}$,

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = \text{log-odds}_1 = log(odds_1)$

$\text{log-odds-ratio}_1 = \frac{\text{log-odds}_1}{\text{log-odds}_0} = log(odds_1) - log(odds_0) = \mathbf{k}\beta_1$

$\rightarrow$ Change: $\text{log-odds-ratio}_1 - \text{log-odds-ratio}_0 = \mathbf{k}\beta_1 - 0 = \mathbf{k}\beta_1$

c. the **odds** will change by a factor (multiplier) of $e^{\mathbf{k}\beta_1}$.

When $X_1 = a$,

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = log\frac{\pi_0}{1-\pi_0} = log(odds_0)$

$odds_0 = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$

When $X_1 = a + \mathbf{k}$,

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = log\frac{\pi_0}{1-\pi_0} = log(odds_0)$

$odds_0 = e^{\beta_0 + \beta_1 (X_1 + \mathbf{k}) + \beta_2 X_2} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathbf{k}\beta_1}$

$\rightarrow$ Change: $odds_1/odds_0 = e^{\mathbf{k}\beta_1}$

d. the **odds** will increase by $(e^{\mathbf{k}\beta_1} - 1) \times 100\%$

$\rightarrow$ Change: $odds_1 - odds_0 = (odds_0 \cdot e^{\mathbf{k}\beta_1}) - odds_0 = odds_0(e^{\mathbf{k}\beta_1} - 1)$

5. Variable selection

a. Methods: $\begin{cases} \text{Stepwise : } \checkmark \\ \text{Best subset: rarely used in GLM} \end{cases}$

b. Criterias

- Significance of βs: Wald's test, LRT
- AIC $= -2log(likelihood(\hat{\beta}_{MLE})) + 2p$
- BIC (SBC) $= -2log(likelihood(\hat{\beta}_{MLE})) + log(n)p$

△ We prefer models with smaller AIC or BIC

△ BIC penalizes number of parameters more than AIC does

6. Predictions

a. $\hat{\pi}$

- Logit:

$$\hat{\pi}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}}}{e^{\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}} + 1}$$

- Probit:

$$\hat{\pi}_i = P\{N(0,1) \leq \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}\}$$

- Complementary log-log:

$$\hat{\pi}_i = 1 - exp\{-exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2})\}$$

b. $\hat{Y}_i$

- If $n_i \neq 1$,
  $\hat{Y}_i = n_i \hat{\pi}_i$
- If $n_i = 1$,
  $\hat{Y}_i = \begin{cases} 1 & , \hat{\pi}_i > c \quad (eg.\ c = 0.5) \\ 0 & , \hat{\pi}_i \leq c \end{cases}$

7. Residuals

a. Pearson residual

- Pearson residual:

$$r_{p_i} = \frac{Y_i - n_i \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

- Standardized (Studentized) Pearson residual:

$$r_{sp_i} = \frac{r_{p_i}}{\sqrt{1 - h_{ii}}}$$

Where $\underbrace{h_{ii}}_{leverage} = diag(H)$

In linear regression:

$$H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

In GLM:

$$H = \hat{\mathbf{W}}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{W}}^{\frac{1}{2}}$$

Where $\hat{\mathbf{W}}$ is the estimated weight matrix:

$$\begin{pmatrix} n_1\hat{\pi}_1(1-\hat{\pi}_1) & 0 & 0 & \dots & 0 \\ 0 & n_2\hat{\pi}_2(1-\hat{\pi}_2) & 0 & \dots & 0 \\ 0 & 0 & n_3\hat{\pi}_3(1-\hat{\pi}_3) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & n_k\hat{\pi}_k(1-\hat{\pi}_k) \end{pmatrix}_{k \times k}$$

b. Deviance residual

- Deviance residual

$$Dev = \underbrace{-2log(Likelihood(\text{Model of interest}))}_{deviance(R)} - \underbrace{(-2log(Likelihood(\text{Saturated model})))}_{deviance(F)}$$

$$= \text{residual deviance} = \sum_i dev_i$$

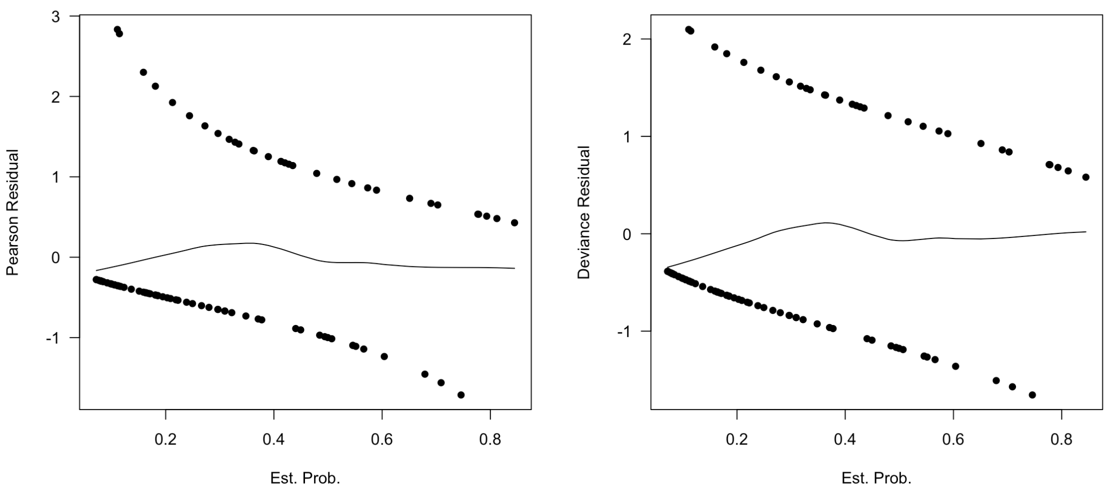
$dev_i =$ the contribution of the i^{th} case to the model deviance

Saturated model	**Model of interest**(eg. logistic model)
$\hat{\pi}_i = \frac{Y_i}{n_i}$	$\hat{\pi}_i = \frac{exp(\hat{\beta_0}+\hat{\beta_1}X_i)}{1+exp(\hat{\beta_0}+\hat{\beta_1}X_i)}$

- Standardized deviance residual

$$\frac{dev_i}{\sqrt{1-h_{ii}}}$$

c. Residual plot



The lowess line is expected to be flat around 0. Patterns or lump indicates an unideal model.

8. Outliers and influential cases

a. Leverage

To identify outlying X observations.
The observation is suspected to be an outlier if:

$$h_{ii} > \frac{2p}{n} \ (\frac{3p}{n})$$

Where $\begin{cases} \text{p: number of parameters} \\ \text{n: sample size} \\ \frac{p}{n} = \bar{h_{ii}} \end{cases}$

b. Cook's distance

To identify influential cases.
(Influential case: with/without this observation, the estimated model changes a lot)
Cook's distance:

$$D_i = \frac{r_{p_i}^2 h_{ii}}{p(1-h_{ii})}$$

measures the influence of the i^{th} observation on the linear procedure.

c. Change in χ^2:

$$\triangle\chi^2_{(i)} = \chi^2 - \chi^2_{(i)} = r^2_{sp_i} = (\text{Standardized Pearson residual})^2$$

d. Change in deviance:

$$\triangle Dev_i = Dev - Dev_i = h_{ii} \cdot r_{sp} + (dev_i)^2$$

9. Goodness of fit

$$\begin{cases} H_0 : g(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 (\text{ model fits data}) \\ H_A : g(E(Y)) \neg \beta_0 + \beta_1 X_1 + \beta_2 X_2 \end{cases}$$

Model: (1) $g()$, (2) $\beta_0 + \beta_1 X_1 + \beta_2 X_2$
Condition: Able to group the covariates (predictors)
Testing procedure:

a. Pearson Goodness of fit test:

$$\begin{aligned}
\chi^2 &= \sum_{i=1}^{I} \frac{(Observed - Expected)^2}{Expected} \\
&= \sum_{i=1}^{I} [\underbrace{\frac{(Y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i}}_{success} + \underbrace{\frac{((n_i - y_i) - n_i(1-\hat{\pi}_i))^2}{n_i(1-\hat{\pi}_i)}}_{failure}] \\
&= \sum_{i=1}^{I} [\frac{Y_i - n_i\hat{\pi}_i}{\sqrt{n_i\pi_i(1\hat{-}\ \hat{\pi})}}] \\
&= \sum_{i=1}^{I} (r_{pi}) = \sum_{i=1}^{I} [(\text{Pearson residuals})^2] \\
&\overset{H_0}{\sim} \chi^2_(df = I - p)
\end{aligned}$$

Where: $\begin{cases} \text{I: number of distinct covariate patterns} \\ \text{p: number of parameters in the model} \end{cases}$

p-value: $P(\chi^2_{(I-p)} > \chi^2)$

b. Deviance test (LRT):

$$\begin{array}{l} \text{Full (saturated model)} : \pi_i \text{ free to change } (\hat{\pi}_i = \frac{Y_i}{n_i}) \\ H_0 \longrightarrow \quad \Big\downarrow \\ \text{Reduced (model of interest)} : log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2) \end{array}$$

$$\begin{aligned} \triangle D = G^2 &= Deviance(R) - Deviance(F) \\ &= \text{(Residual) Deviance of the model of interest} - 0 \\ &= \sum_{i=1}^{I} (dev_i^2) \\ &\overset{H_0}{\sim} \chi^2_{(I-p)} \end{aligned}$$

Notice that: $a. \overset{n\to\infty}{=} b.$

c. Hosmer-Lemeshow test ($Y_i \sim Bernoulli(\pi_i)$)

(1) Fit the model, compute $\hat{\pi}_i$

(2) Group observations by $\hat{\pi}_i$

- Option 1: According to $\hat{\pi}_i$ divide all n observations into k groups of equal/similar size
- Option 2: Divide $\hat{\pi}_i$ into k equal fractions, and regard all observations in a fraction as a group

(3) Within each group,

$$\begin{cases} \text{Observed success}: & \sum_{Y_i \in group_k} Y_i \\ \text{Observed failure}: & \sum_{Y_i \in group_k} (1 - Y_i) \\ \text{Expected success}: & \sum_{Y_i \in group_k} \hat{\pi}_i \\ \text{Expected failure}: & \sum_{Y_i \in group_k} (1 - \hat{\pi}_i) \end{cases}$$

(4) Pearson Chi-square:

$$\sum (\frac{(Observed - Expected)^2}{Expected}) \overset{H_0}{\sim} \chi^2_{(df=k-p)}$$

10. Receiver Operation Charateristic Curve (ROC Curve)

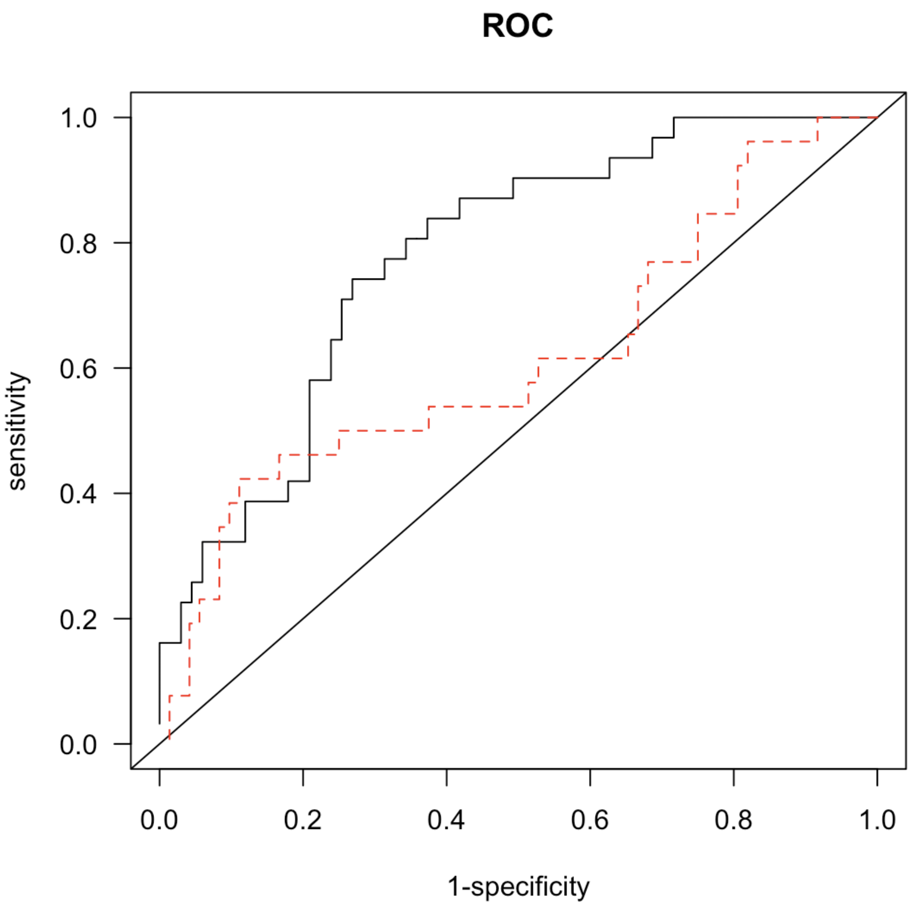
- Cutoff: c

  Prediction: $\hat{Y_i} = \begin{cases} 1, & if \hat{\pi}_i > c \\ 0, & if \hat{\pi}_i \leq c \end{cases}$

- Terms:

$$\begin{cases} \text{Specificity (TNR)} = P(\hat{Y}=0|Y=0) = P(\text{correctly classify } \hat{Y}_i \text{ as } 0) \\ \text{Sensitivity (TPR)} = P(\hat{Y}=1|Y=1) = P(\text{correctly classify } \hat{Y}_i \text{ as } 1) \\ \text{FPR (False Negative Rate)} = P(\hat{Y}=0|Y=1) = 1 - \text{Sensitivity} \\ \text{FNR (False Positive Rate)} = P(\hat{Y}=1|Y=0) = 1 - \text{Specificity} \end{cases}$$

- Graph



  - Cutoff: choose a cutoff to keep both FNR and FPR relatively low. or, to minimize overall error rate.
  - AUR (Area Under ROC curve): large area indicates a good model.
  - ROC as validation tool: if ROC_{train} and ROC_{test} are similar, we are more confident about the model.