Assessing the model

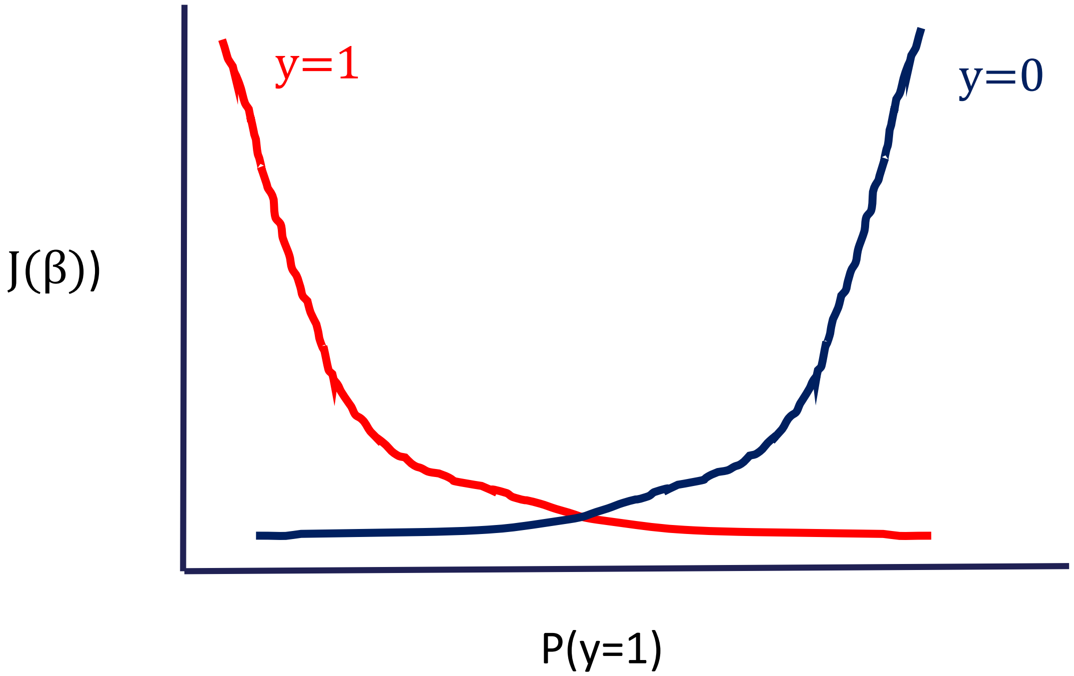
Logistic Regression Model

$$P(y=1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

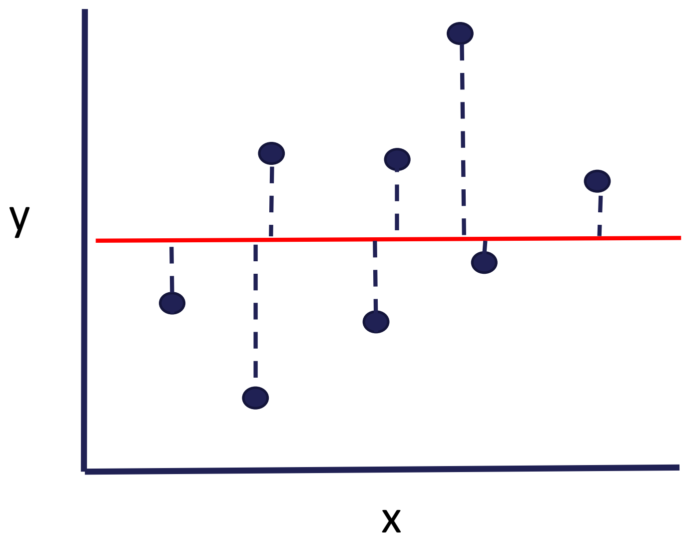
- X are the independent variables.
- Y is the target, which is class target.

Log-loss

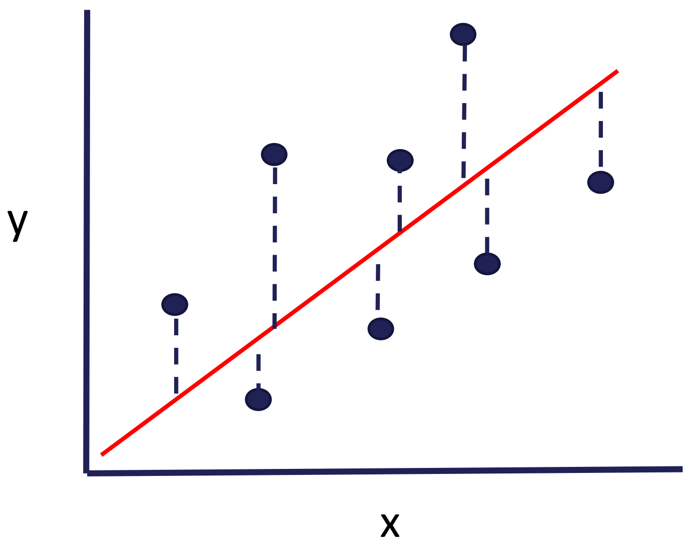
$$J(\beta) = -\frac{1}{m}\sum y \log(P(y=1)) + (1-y) \log(1 - P(y=1))$$



Goodness of fit – linear model



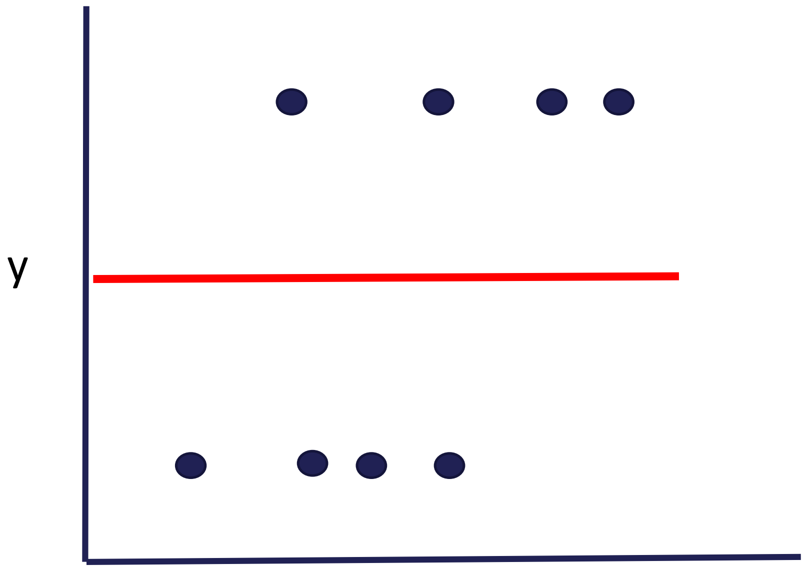
$Predict\ target\ mean$



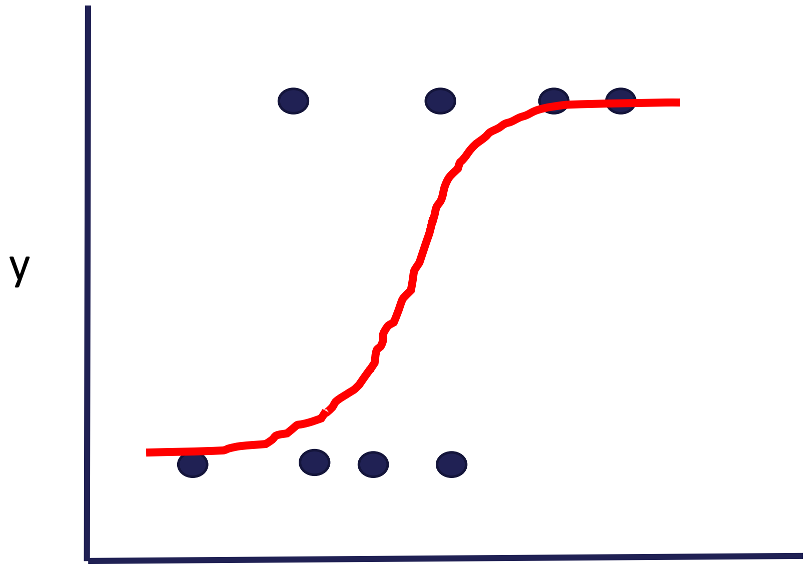
$Model$

$$Metric = Baseline\ deviation - Left\ over\ deviation$$

Goodness of fit- logistic model



$Predict\ target\ mean$



$Model$

$$Metric = Baseline\ deviation - Left\ over\ deviation$$

Log-loss

$$J(\beta) = -\frac{1}{m}\sum y\,\log(P(y=1)) + (1-y)\,\log\,(1-P(y=1))$$

Log-loss

$$J(\beta) = -\frac{1}{m}\sum y \log(P(y=1)) + (1-y) \log(1- P(y=1))$$

Target	Baseline
0	0.6
0	0.6
1	0.6
1	0.6
1	0.6

Log-loss

$$J(\beta) = -\frac{1}{m}\sum y \log(P(y=1)) + (1-y)\log(1-P(y=1))$$

Target	Baseline
0	0.6
0	0.6
1	0.6
1	0.6
1	0.6

-1/5 * (log(0.6) + log(0.6) + log(0.6) + log(1-0.6) + log(1-0.6)) =

-1/5 * (-0.51 – 0.51 – 0.51 – 0.91 – 0.91) =

0.67

Log-loss

$$J(\beta) = -\frac{1}{m}\sum y \log(P(y=1)) + (1-y) \log(1- P(y=1))$$

Target	Logit
0	0.2
0	0.3
1	0.6
1	0.7
1	0.8

-1/5 * (log(0.6) + log(0.7) + log(0.8) + log(1-0.2) + log(1-0.3)) =

-1/5 * (-0.51 – 0.35 – 0.22 -0.22 – 0.35) =

0.33

Log-loss

$$J(\beta) = -\frac{1}{m}\sum y \log(P(y=1)) + (1-y) \log(1 - P(y=1))$$

The smaller the log-loss ➔ the better the model

Log-likelihood

$$J(\beta) = -\frac{1}{m}\sum \underbrace{y \log(P(y=1)) + (1-y) \log (1- P(y=1))}_{\text{log-likelihood}}$$

Log-likelihood

$$LL = \sum y \log(P(y=1)) + (1-y) \log(1- P(y=1))$$

Target	Base	Logit
0	0.6	0.2
0	0.6	0.3
1	0.6	0.6
1	0.6	0.7
1	0.6	0.8

- LL_base = - 3.36
- LL_logit = -1.61

Log-likelihood

$$LL = \sum y \log(P(y{=}1)) + (1-y)\log(1-P(y{=}1))$$

Target	Base	Logit
0	0.6	0.2
0	0.6	0.3
1	0.6	0.6
1	0.6	0.7
1	0.6	0.8

- LL_base = - 3.36
- LL_logit = -1.61

✓ Indicates how much information is left unexplained.

✓ Larger LL ➔ poorer fit

Deviance

$$LL = \sum y \log(P(y=1)) + (1-y) \log(1- P(y=1))$$

$$\text{Deviance} = -2LL$$

Deviance

$$LL = \sum y \log(P(y=1)) + (1-y) \log (1 - P(y=1))$$

$$\text{Deviance} = -2LL$$

Target	Base	Logit
0	0.6	0.2
0	0.6	0.3
1	0.6	0.6
1	0.6	0.7
1	0.6	0.8

- -2LL_base = -3.36 * (-2) = 6.73
- -2LL_logit = -1.61 * (-2) = 3.34

Deviance

$$LL = \sum y \log(P(y{=}1)) + (1-y) \log(1 - P(y{=}1))$$

Target	Base	Logit
0	0.6	0.2
0	0.6	0.3
1	0.6	0.6
1	0.6	0.7
1	0.6	0.8

Deviance= -2LL

- -2LL_base = -3.36 * (-2) = 6.73
- -2LL_logit = -1.61 * (-2) = 3.34

The smaller the deviance, the better the fit.

Goodness of fit

$$x^2 = \text{Deviance(baseline)} - \text{Deviance(model)}$$

- If deviance is large ➔ model is poor
- Baseline ➔ largest deviance
- If the model is good, its deviance is small ➔ X^2 is large
- If the model is bad, its deviance is large ➔ X^2 is small

Goodness of fit

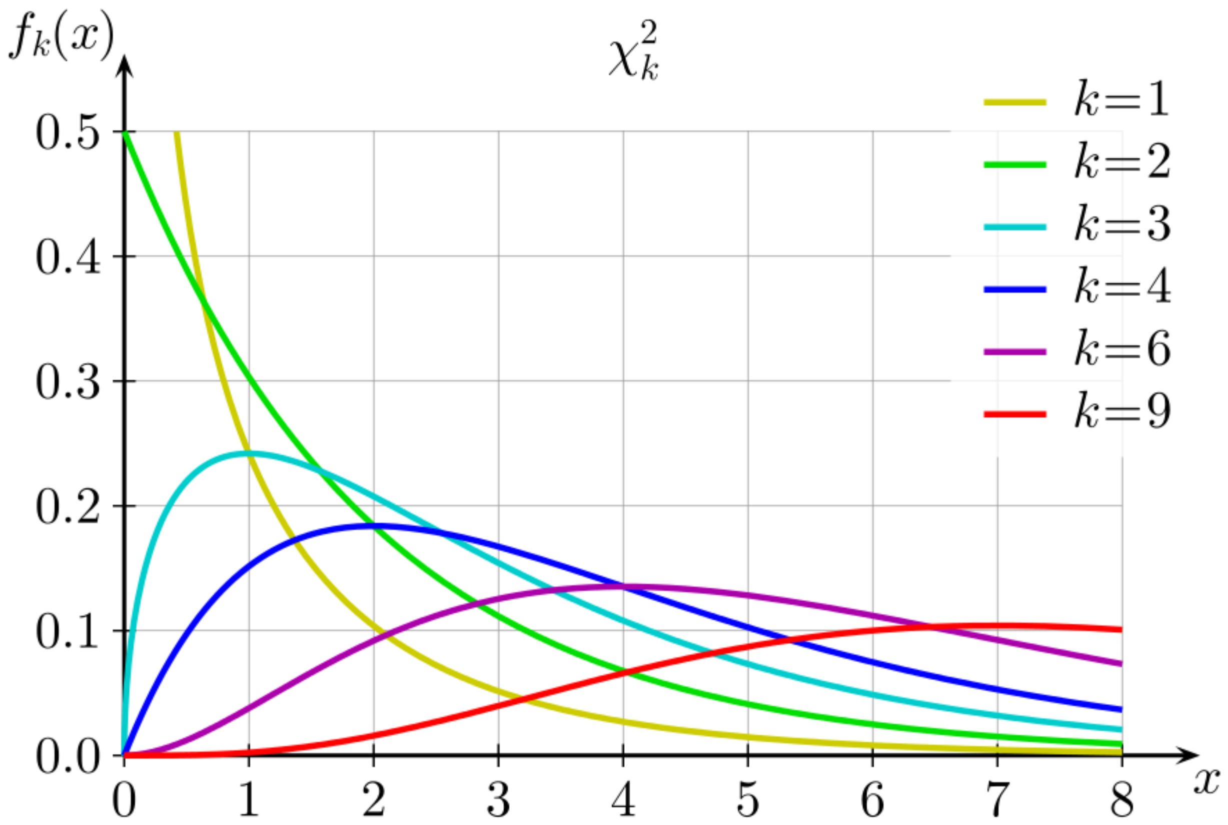
$$x^2 = \text{Deviance(baseline)} - \text{Deviance(model)}$$

- x^2 follows a chi-squared distribution.
- Chi-squared tests the null hypothesis: The logistic regression is not better than the baseline.
- Degrees of freedom: number of predictors.

Chi-squared distribution

Large x^2 values or small p-values

➔ the logistic regression is better than the baseline.



R-statistic

$$R^2 = \frac{\text{Deviance(baseline)} - \text{Deviance(logistic regression)}}{\text{Deviance(baseline)}}$$

- Baseline ➔ largest deviance
- If the model is as bad as the baseline ➔ R^2 is close to 0.
- If the model is good, its deviance is small ➔ R^2 is closer to 1.
- The smallest the deviance of the model, the greater the R^2.

THANK YOU

www.trainindata.com