Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

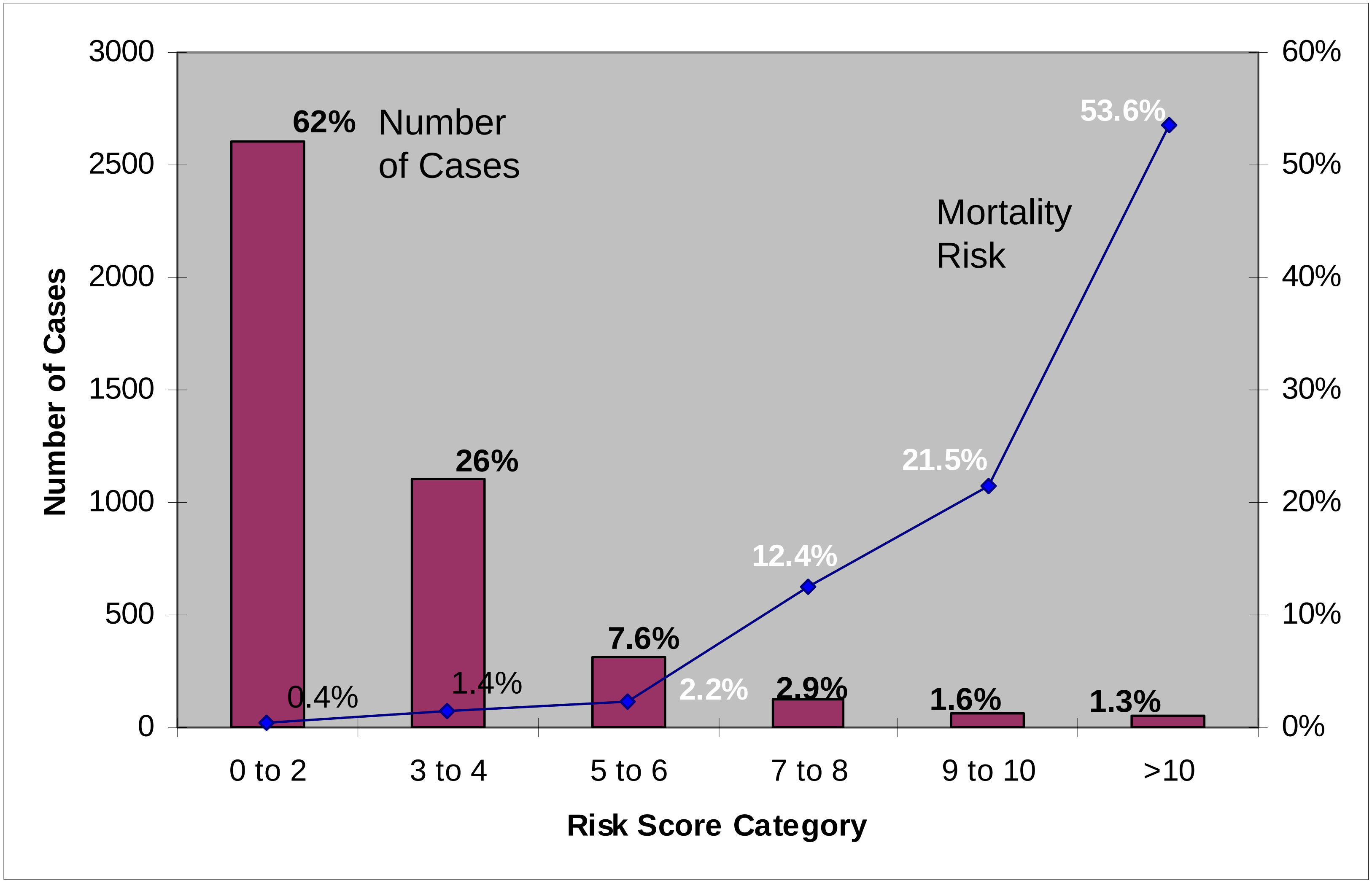
6.873/HST.951 Medical Decision Support
Fall 2005

Logistic Regression
Maximum Likelihood Estimation

Lucila Ohno-Machado

Risk Score of Death from Angioplasty

Unadjusted Overall Mortality Rate = 2.1%

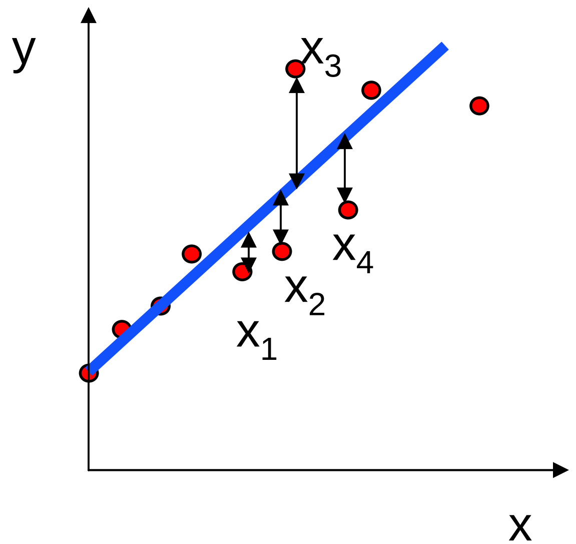


Linear Regression
Ordinary Least Squares (OLS)

Minimize Sum of Squared Errors (*SSE*)

n data points

i is the subscript for each point



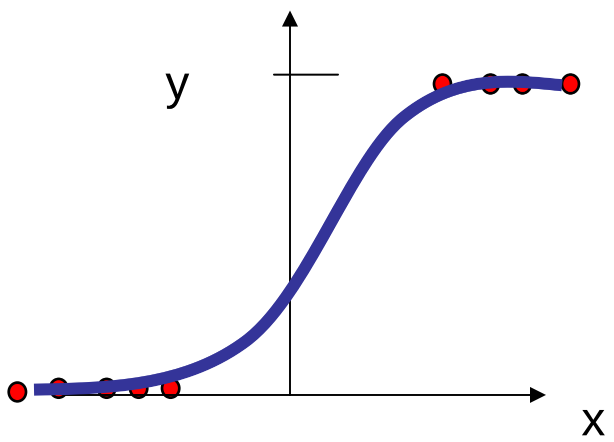
$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

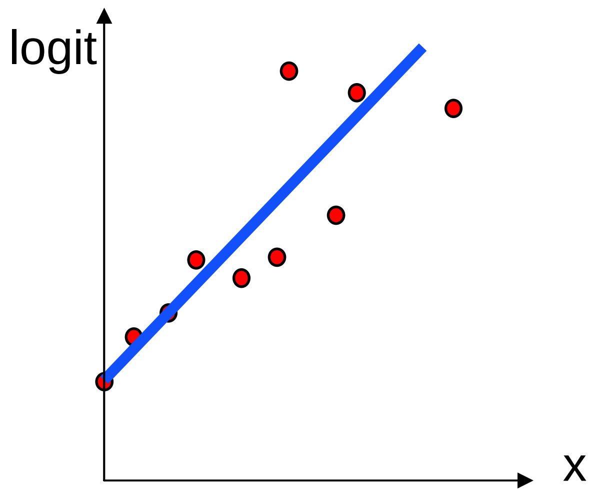
Logit

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

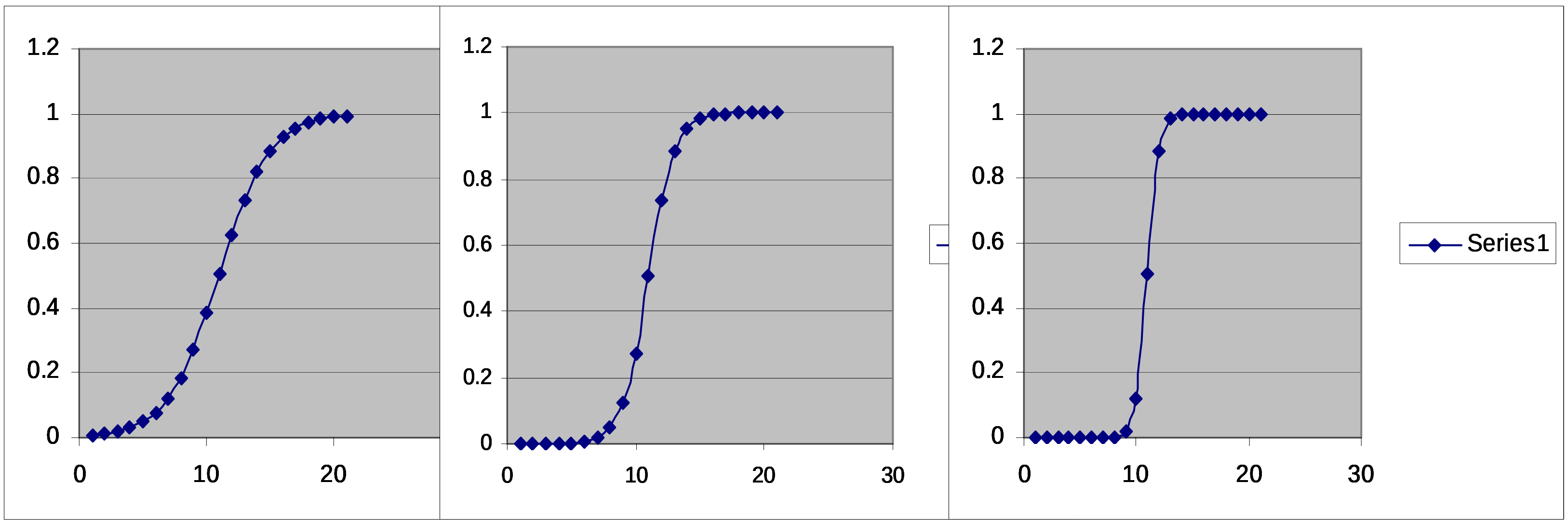
$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{e^{\beta_0+\beta_1 x_i}+1}$$



$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 1+\beta_1 x_i$$



Increasing β



Finding β_0

- Baseline case

$$p_i = \frac{1}{1+e^{-(\beta_0)}}$$

	Blue(1)	Green(0)	
Death	28	22	50
Life	45	52	97
Total	73	74	147

$$0.297 = \frac{1}{1+e^{-(\beta_0)}}$$

$$\beta_0 = -0.8616$$

Odds ratio

- Odds: p/(1-p)
- Odds-ratio

$$OR = \frac{\dfrac{p_{death|blue}}{1-p_{death|blue}}}{\dfrac{p_{death|green}}{1-p_{death|green}}}$$

	Blue	Green	
Death	28	22	50
Life	45	52	97
Total	73	74	147

$$OR = \frac{28/45}{22/52} = 1.47$$

What do coefficients mean?

$e^{\beta_{color}} = OR_{color}$

	Blue	Green	
Death	28	22	50
Life	45	52	97
Total	73	74	147

$$OR = \frac{28/45}{22/52} = 1.47$$

$$e^{\beta_{color}} = 1.47$$

$$\beta_{color} = 0.385$$

$$p_{blue} = \frac{1}{1+e^{-(-0.8616+0.385)}} = 0.383$$

$$p_{green} = \frac{1}{1+e^{0.8616}} = 0.297$$

What do coefficients mean?

$$e^{\beta_{age}} = OR_{age}$$

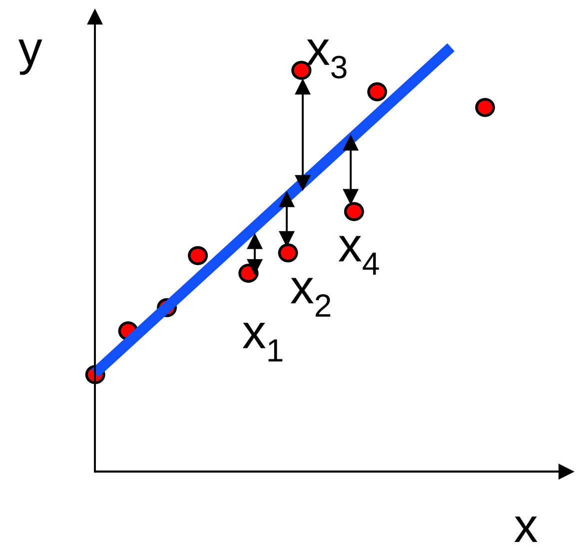
	Age49	Age50	
Death	28	22	50
Life	45	52	97
Total	73	74	147

$$OR = \frac{\dfrac{p_{death|age=50}}{1 - p_{death|age=50}}}{\dfrac{p_{death|age=49}}{1 - p_{death|age=49}}}$$

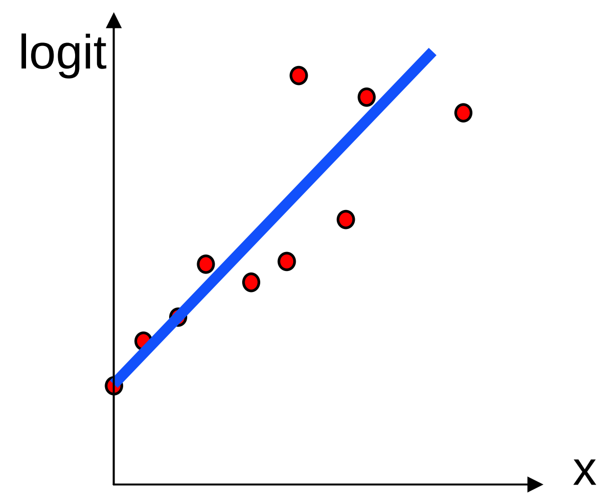
Why not search using OLS?

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



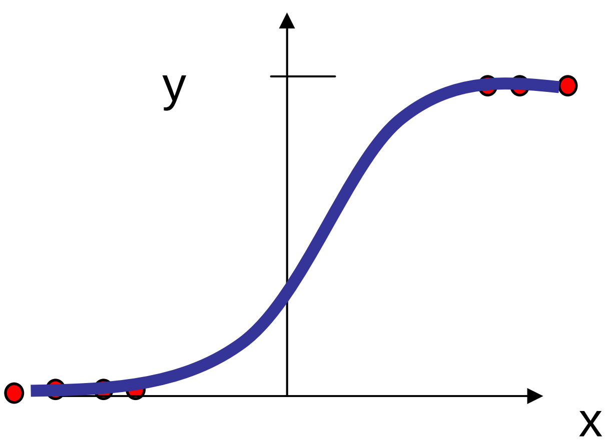
$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 1 + \beta_1 x_i$$

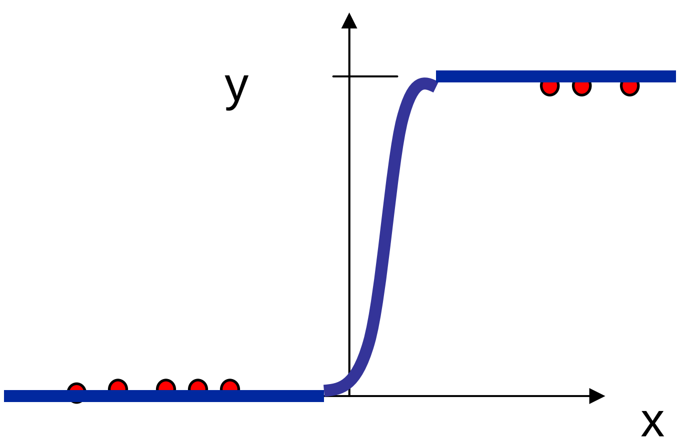


P(model | data) ?

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$$

If only intercept is allowed, which value would it have?





x

P (data | model) ?

P(data|model) = [P(model | data) P(data)] / P(model)

When comparing models:

P(model): assume all the same (ie, chances of being a model with high coefficients the same as low, etc)

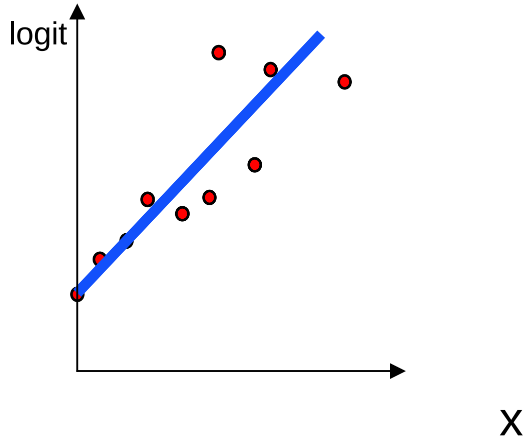
P(data): assume it is the same

Then,

P(data | model) α P(model | data)

Maximum Likelihood Estimation

- Maximize P(data | model)
- Maximize the probability that we would observe what we observed (given assumption of a particular model)
- Choose the best parameters from the particular model



Maximum Likelihood Estimation

- Steps:
  - Define expression for the probability of data as a function of the parameters
  - Find the values of the parameters that maximize this expression

Likelihood Function

$$L = \Pr(Y)$$

$$L = \Pr(y_1, y_2, \ldots, y_n)$$

$$L = \Pr(y_1)\Pr(y_2)\ldots\Pr(y_n) = \prod_{i=1}^{n} \Pr(y_i)$$

Likelihood Function
Binomial

$$L = \Pr(Y)$$

$$L = \Pr(y_1, y_2, \ldots, y_n)$$

$$L = \Pr(y_1)\Pr(y_2)\ldots\Pr(y_n) = \prod_{i=1}^{n} \Pr(y_i)$$

$$\Pr(y_i = 1) = p_i$$

$$\Pr(y_i = 0) = (1 - p_i)$$

$$\Pr(y_i) = p_i^{\,y_i}(1 - p_i)^{1-y_i}$$

Likelihood Function

$$L = \prod_{i=1}^{n} \Pr(y_i) = \prod_{i=1}^{n} p_i^{\,y_i} (1 - p_i)^{1-y_i}$$

$$L = \prod_{i=1}^{n} \left(\frac{p_i}{(1 - p_i)} \right)^{y_i} (1 - p_i)$$

Log Likelihood Function

$$L = \prod_{i=1}^{n} \Pr(y_i) = \prod_{i=1}^{n} p_i^{\,y_i} (1 - p_i)^{1-y_i}$$

$$L = \prod_{i=1}^{n} \left(\frac{p_i}{(1 - p_i)} \right)^{y_i} (1 - p_i)$$

$$\log L = \sum_i y_i \log\left(\frac{p_i}{(1 - p_i)} \right) + \sum_i \log(1 - p_i)$$

Log Likelihood Function

$$\log L = \sum_i y_i \log\left(\frac{p_i}{(1-p_i)}\right) + \sum_i \log(1-p_i)$$

$$\log L = \sum_i y_i (\beta x_i) - \sum_i \log(1+e^{\beta x_i})$$

Since model is the logit

Maximize

$$\log L = \sum_i y_i(\beta x_i) - \sum_i \log(1 + e^{\beta x_i})$$

Maximize

$$\log L = \sum_i y_i(\beta x_i) - \sum_i \log(1 + e^{\beta x_i})$$

$$\frac{\partial \log L}{\partial \beta} = \sum_i y_i x_i - \sum_i \hat{y}_i x_i = 0$$

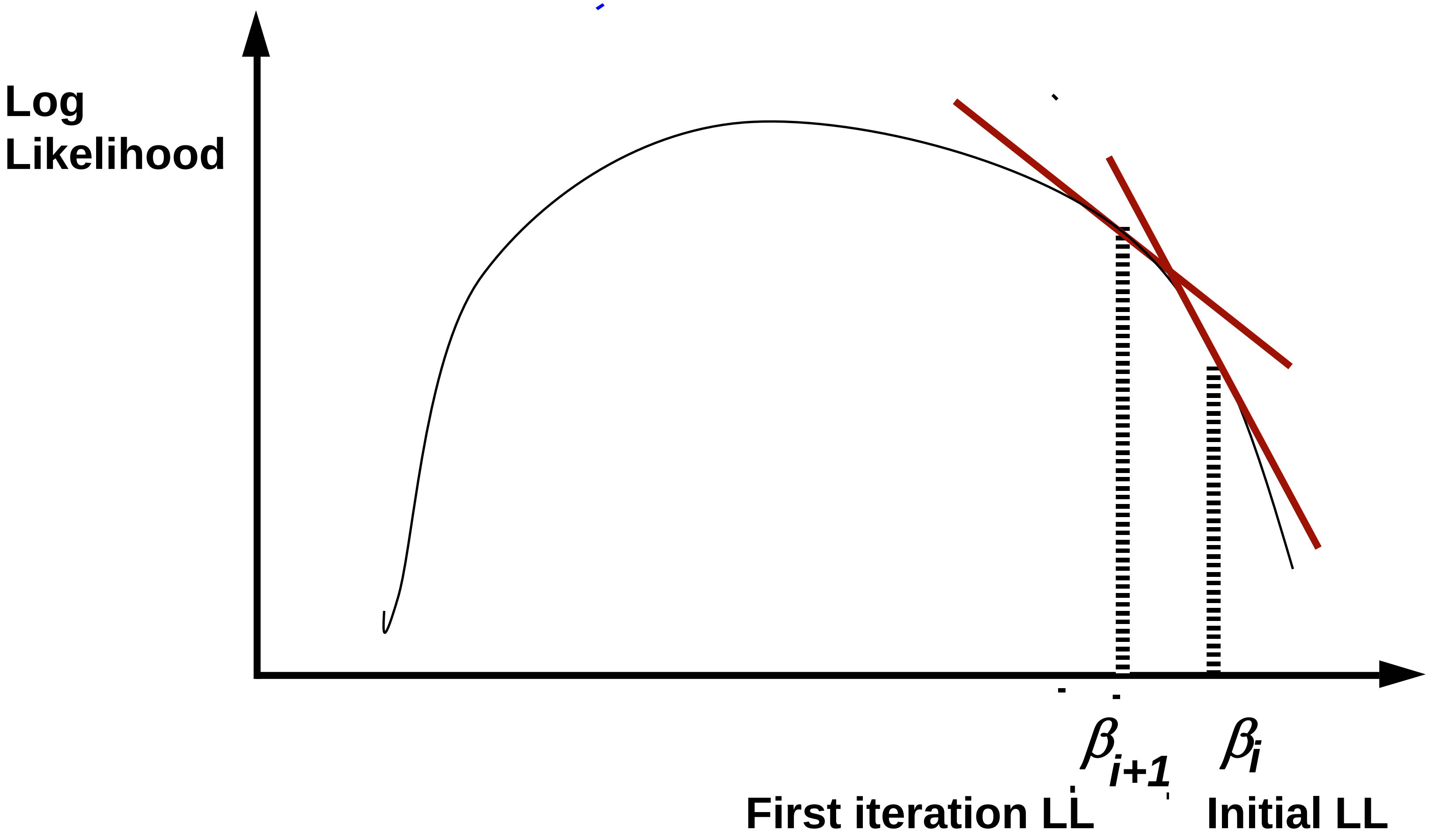
$$\hat{y}_i = \frac{1}{1 + e^{-\beta x_i}}$$

Not easy to solve because y-hat is non-linear, need to use iterative methods: most popular is Newton-Raphson

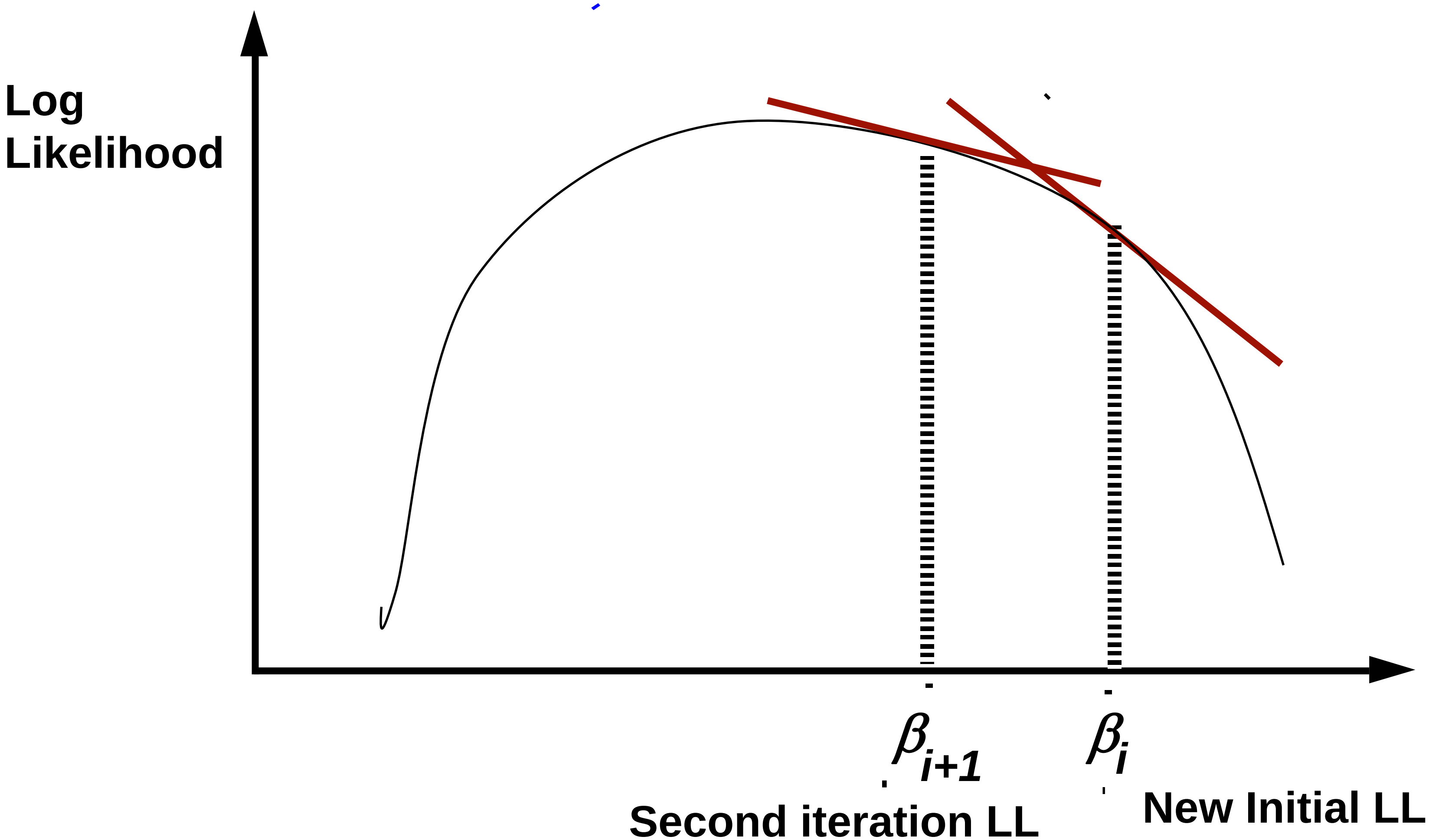
Newton-Raphson

- Start with random or zero βs
- “walk” in the “direction” that maximizes MLE
  - how big a step (Gradient or Score)
  - direction

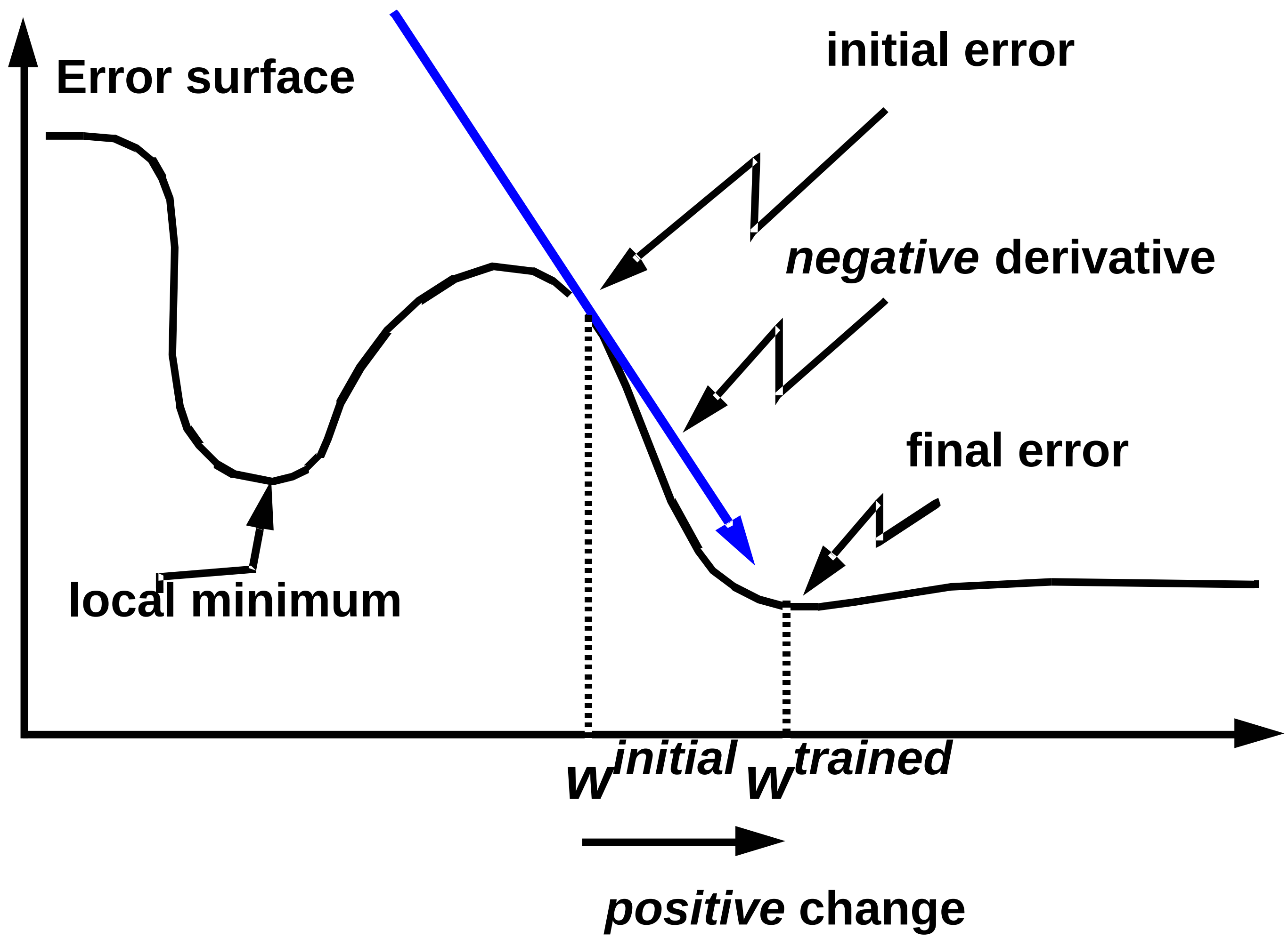
Maximizing the LogLikelihood



Maximizing the LogLikelihood



Similar iterative method to Minimizing the Error in Gradient Descent (neural nets)



Newton-Raphson Algorithm

$$\log L = \sum_i y_i(\beta x_i) - \sum_i \log(1 + e^{\beta x_i})$$

$$U(\beta) = \frac{\partial \log L}{\partial \beta} = \sum_i y_i x_i - \sum_i \hat{y}_i x_i$$ Gradient

$$I(\beta) = \frac{\partial^2 \log L}{\partial \beta \, \partial \beta'} = -\sum_i x_i x_i' \hat{y}_i (1 - \hat{y}_i)$$ Hessian

$$\beta_{j+1} = \beta_j - I^{-1}(\beta_j) U(\beta_j)$$ a step

Convergence

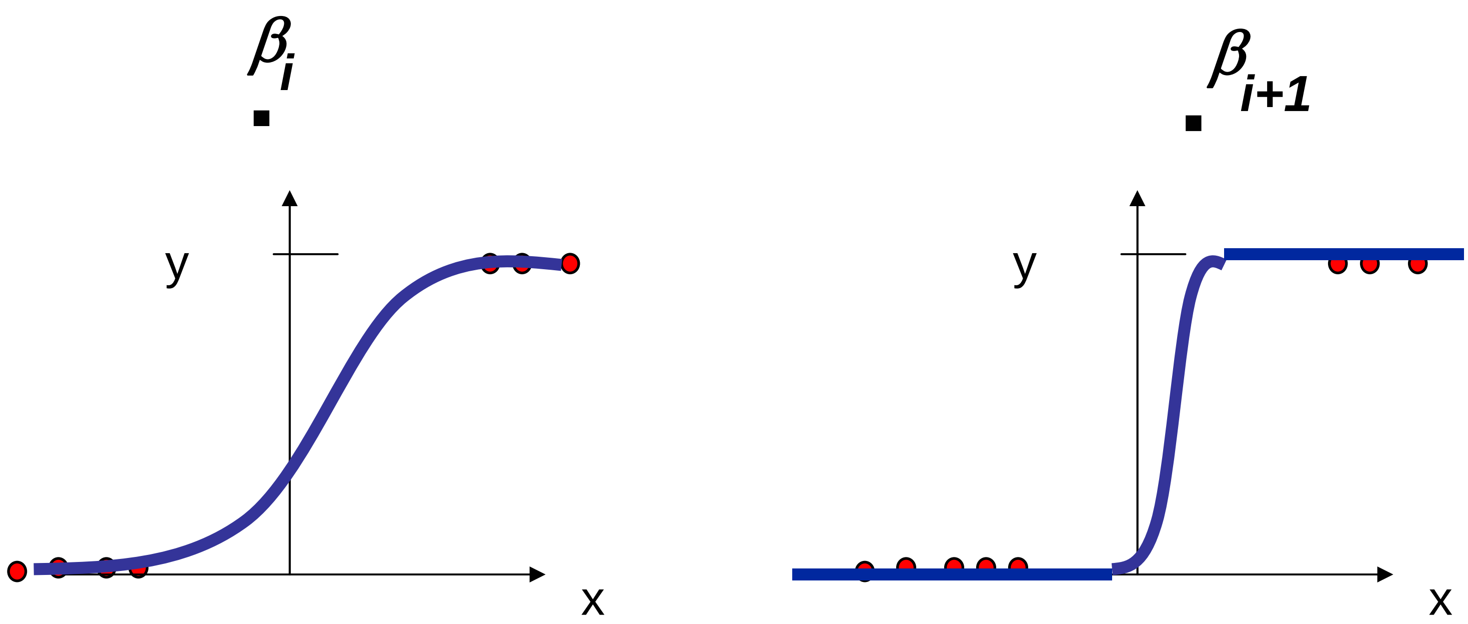
- Criterion

$$\left| \frac{\beta_{j+1} - \beta_j}{\beta_j} \right| < .0001$$

- Convergence problems: complete and quasi-complete separation

Complete separation

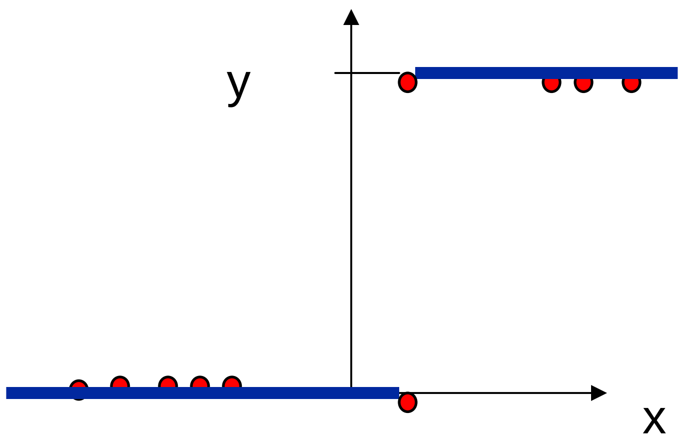
MLE does not exist (ie, it is infinite)



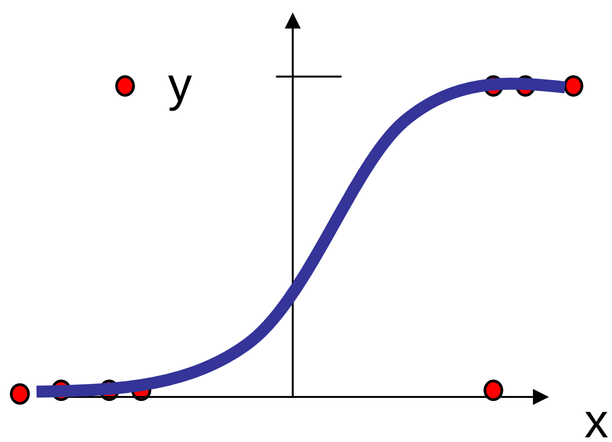
x

Quasi-complete separation

Same values for predictors, different outcomes



No (quasi)complete separation is fine to find MLE



How good is the model?

- Is it better than predicting the same prior probability for everyone? (ie, model with just β_0)
- How well do the training data fit?
- How well does is generalize?

Generalized likelihood-ratio test

- Are β_1, β_2, …, β_n different from 0?

$$L = \prod_{i=1}^{n} \Pr(y_i) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

$$\log L = \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

$$G = -2 \log L_o + 2 \log L_1$$

G has χ^2 distribution

$$cross-entropy\_error = -\sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

AIC, SC, BIC

- To compare models
- Akaike's Information Criterion, *k* parameters

$$AIC = -2\log L + 2k$$

- Schwartz Criterion, Bayesian Information Criterion, n cases

$$BIC = -2\log L + k\log n$$

Summary

- Maximum Likelihood Estimation is used in finding parameters for models
- MLE maximizes the probability that the data obtained would have been generated by the model
- Coming up: goodness-of-fit (how good are the predictions?)
  - How well do the training data fit?
  - How well does is generalize?