DSS
Decision Sciences & Systems

Technische Universität München

Tutorial Business Analytics

Tutorial 4: Generalized Linear Models

Decision Sciences & Systems (DSS)
Department of Informatics
TU München

Tutorial Business Analytics

Agenda

1. Generalized Linear Models
2. Logistic Regression
3. Poisson Regression
4. Maximum Likelihood Estimation
5. Evaluation and Goodness-of-Fit

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Generalized Linear Models

- GLMs are a general class of linear models
- Consist of three components:

- **Random:** Identifies dependent variable μ and probability distribution
- **Systematic:** Identifies the set of explanatory variables $(X_1, \dots, X_k)$
- **Link function:** Identifies function of μ that is linear

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Example: Linear regression uses identity link ($g(\mu) = \mu$)

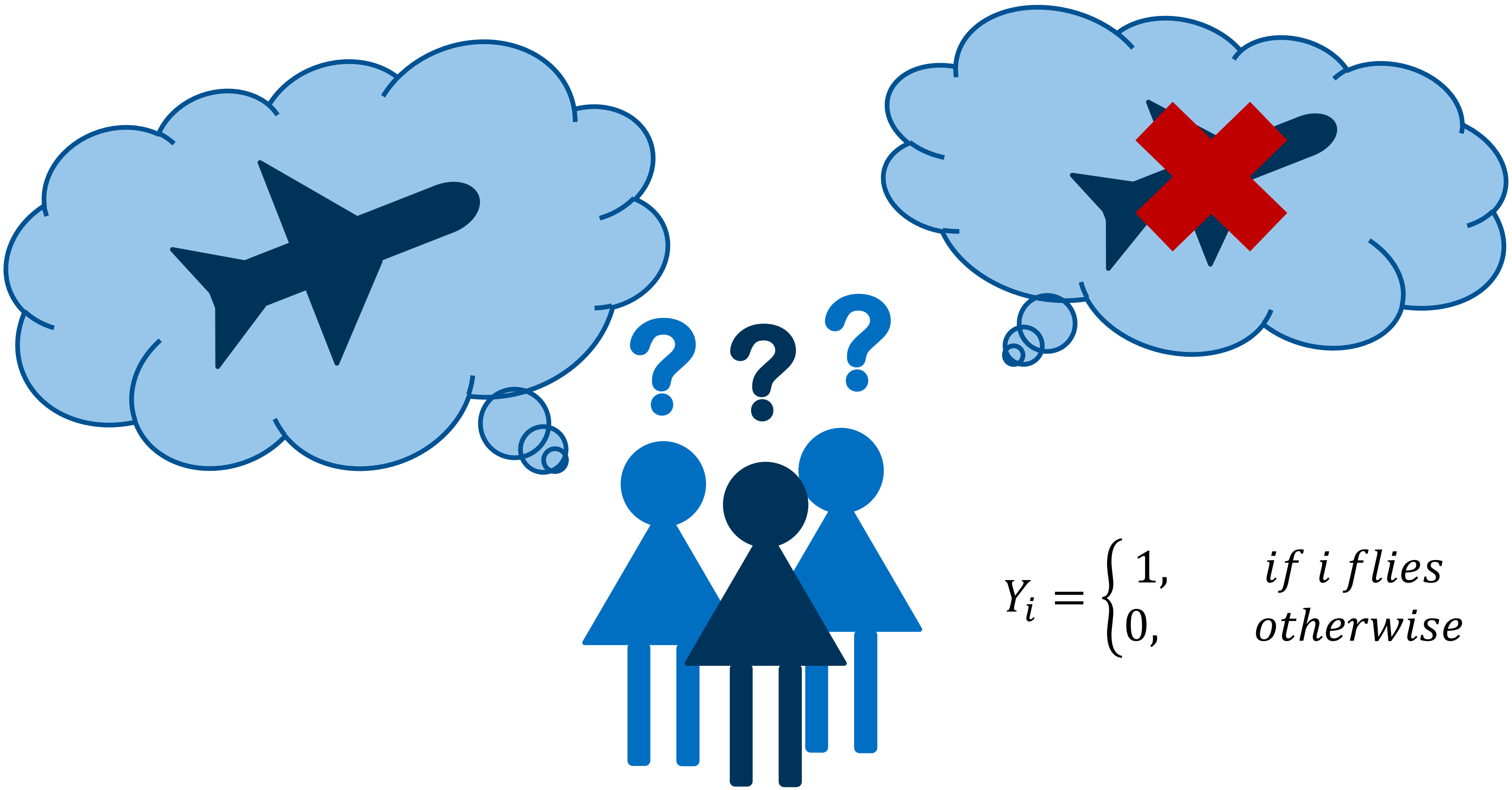
Question: Which link function could be useful for a binary dependent variable?

© Chair of Decision Sciences and Systems,
Technical University of Munich

DSS
Decision Sciences & Systems

Tutorial Business Analytics

Logistic Regression - Motivation

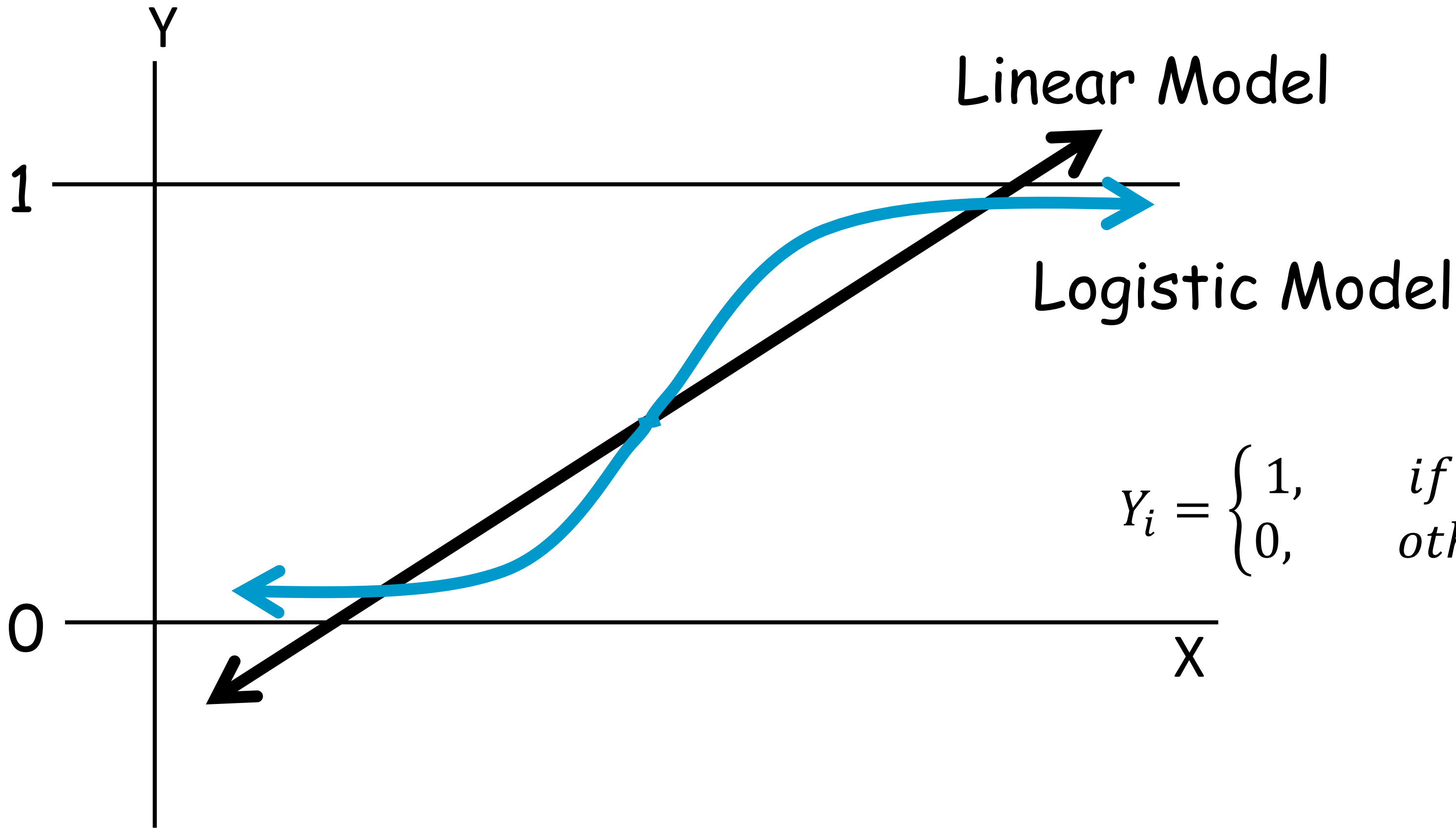


$$Y_i = \begin{cases} 1, & if\ i\ flies \\ 0, & otherwise \end{cases}$$

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Logistic Regression - Motivation



$$Y_i = \begin{cases} 1, & if\ i\ flies \\ 0, & otherwise \end{cases}$$

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

From Logistic Function to Logit

Logistic Function: $p(x_i) = \frac{e^{x_i{}'\beta}}{1+e^{x_i{}'\beta}}$

transform …

Logit: $\ln(\frac{p(x_i)}{1-p(x_i)}) = x_i{}'\beta$

$\Leftrightarrow$ $\boldsymbol{\frac{p(x_i)}{1-p(x_i)}} = e^{x_i{}'\beta}$ **odds**

$\boxed{\text{Logistic Regression:} \quad \ln(\frac{p(x_i)}{1-p(x_i)}) = x_i{}'\beta + \varepsilon_i}$

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Interpreting the coefficient of logistic regression

$x_{ij} \in x_i$: $\qquad \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = x_i{}'\beta$

$(x_{ij}+1) \in \tilde{x}_i$: $\qquad \ln\left(\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}\right) = \tilde{x}_i{}'\beta$

$$\ln\left(\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}\right) - \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \tilde{x}_i{}'\beta - x_i{}'\beta = \beta_j$$

$$\Leftrightarrow \qquad \beta_j = \ln\left(\frac{\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}}{\frac{p(x)}{1-p(x)}}\right)$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}}{\frac{p(x_i)}{1-p(x_i)}}$$

odds ratio

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Summary: Interpreting the coefficient of logistic regression

Effect of change in x_{ij}: on **log-odds (A)**, **odds (B)** and **probability (C)**

$$\Delta x_{ij} = 1 > 0$$

$$\Rightarrow \quad \Delta\, ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \ln\left(\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}\right) - \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_j \qquad \textbf{(A)}$$

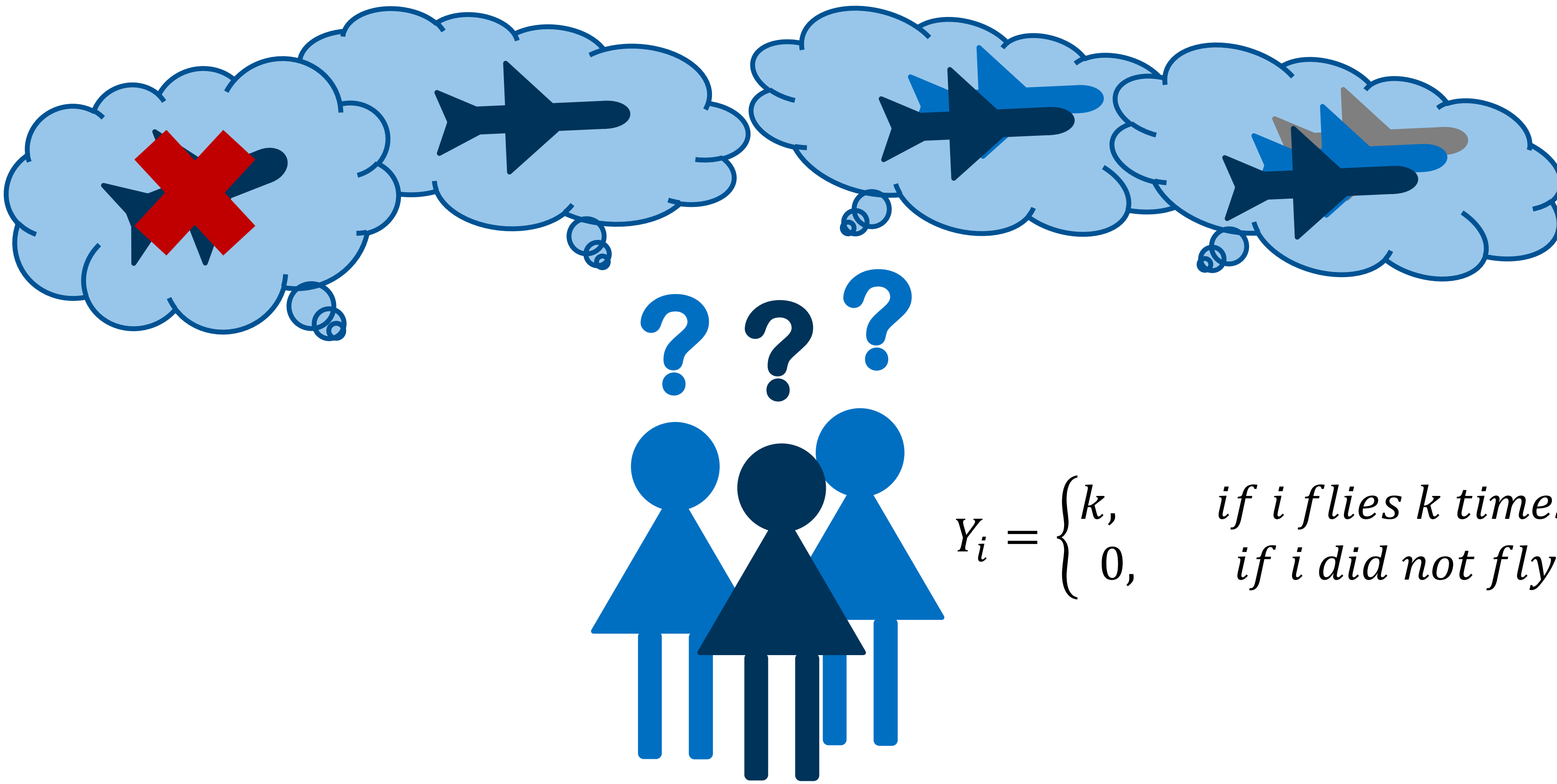
$$\Leftrightarrow \quad e^{\beta_j} = \frac{\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}}{\frac{p(x_i)}{1-p(x_i)}} \qquad \textbf{(B), (C)}$$

$\boldsymbol{\beta_j}$	$\boldsymbol{ln(\frac{p}{1-p})}$ **(A)**	$\boldsymbol{\frac{p}{1-p}}$ **(B)**	$\boldsymbol{p}$ **(C)**
$\beta_j > 0$	increases by β_j	increases by a factor of e^{β_j}	Magnitude of increase unknown
$\beta_j < 0$	decreases by β_j	decreases by a factor of e^{β_j}	Magnitude of decrease unknown

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Poisson Regression – Motivation



$$Y_i = \begin{cases} k, & if\ i\ flies\ k\ times \\ 0, & if\ i\ did\ not\ fly \end{cases}$$

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

From Incidence Rate to Link Function

Incidence Rate: $\mu(x) = e^{x_i{}'\beta}$

transform …

Link Function: $\ln\big(\mu(x)\big) = x_i{}'\beta$

Poisson Regression: $\boxed{\ln\big(\mu(x)\big) = x_i{}'\beta + \varepsilon_i}$

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Interpreting the coefficient of poisson regression

$x_{ij} \epsilon x_i$: $\qquad\qquad \ln\big(\mu(x_i)\big) = x_i{}'\beta$

$\big(x_{ij} + 1\big) \epsilon \tilde{x}_i$: $\qquad\qquad \ln\big(\mu(\tilde{x}_i)\big) = \tilde{x}_i{}'\beta$

$$\ln\big(\mu(\tilde{x}_i)\big) - \ln\big(\mu(x_i)\big) = \tilde{x}_i{}'\beta - x_i{}'\beta = \beta_j$$

$$\Leftrightarrow \qquad \beta_j = \ln\left(\frac{\mu(\tilde{x}_i)}{\mu(x_i)}\right)$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\mu(\tilde{x}_i)}{\mu(x_i)}$$

incidence rate ratio

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Summary: Interpreting the coefficient of poisson regression

Effect of change in x_{ij}: on **log-incidence rate (A)**, **incidence rate (B)**

$$\Delta x_{ij} = 1 > 0$$

$$\Rightarrow \qquad \Delta \ln\big(\mu(x_i)\big) = \ln\big(\mu(\tilde{x}_i)\big) - \ln\big(\mu(x_i)\big) = \beta_j \qquad \textbf{(A)}$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\mu(\tilde{x}_i)}{\mu(x_i)} \qquad \textbf{(B)}$$

$\boldsymbol{\beta_j}$	$\boldsymbol{ln\big(\mu(x_i)\big)}$ **(A)**	$\boldsymbol{\mu(x_i)}$ **(B)**
$\beta_j > 0$	increases by β_j	**increases by a factor of** e^{β_j}
$\beta_j < 0$	decreases by β_j	**decreases by a factor of** e^{β_j}

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Maximum Likelihood Estimation

Goal: Maximize the joint probability of observing the set of dependent variables of the random sample

- Logistic regression: $L = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$ with $p = \frac{e^{X\beta}}{1+e^{X\beta}}$
- Poisson regression: $L = \prod_{i=1}^{n} p$ with $p = \frac{e^{X\beta y}}{y!} e^{-e^{X\beta}}$

Use numerical algorithm to find the maximum → gradient ascent

$k = 1$, feasible start point $\beta^{(1)} \in \mathbb{R}^n$, parameter $\varepsilon > 0$

While ($\left\|\nabla L(\beta^{(k)})\right\| \geq \varepsilon$) {

- Choose step size $\alpha > 0$
- Set $\beta^{(k+1)} = \beta^{(k)} + \alpha^* \nabla L\left(\beta^{(k)}\right)$
- $k++$

}

© Chair of Decision Sciences and Systems,
Technical University of Munich

Tutorial Business Analytics

Evaluation and Goodness-of-Fit

- Null deviance: $-2\ln\left(\mathrm{L}(\mathrm{null})\right)$
- Residual deviance: $-2\ln\left(\mathrm{L}(\mathrm{fitted})\right)$
- McFadden R²:

$$R^2_{McFadden} = 1 - \frac{LL(fitted)}{LL(null)}$$

- Likelihood ratio test: Does fitted model explain significantly more variance than null model?

$$\mathrm{D} = -2\ln\left(\frac{\mathrm{L}(\mathrm{null})}{\mathrm{L}(\mathrm{fitted})}\right) = -2\left(\mathrm{LL}(\mathrm{null}) - \mathrm{LL}(\mathrm{fitted})\right)$$

- Wald test: Is a particular coefficient significant?

$$\mathrm{H}_0\text{: } \beta_i = 0$$

© Chair of Decision Sciences and Systems,
Technical University of Munich

DSS
Decision Sciences & Systems

Tutorial Business Analytics

Exemplary R Output

```
> model <- glm( diabetes ~ glucose + mass + age, data = diabData, family = binomial)
> summary(model)

Call:
glm(formula = diabetes ~ glucose + mass + age, family = binomial,
    data = diabData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6030  -0.6666  -0.3815   0.6765   2.4804

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.677346   1.041873  -9.288  < 2e-16 ***
glucose      0.036266   0.004906   7.391 1.45e-13 ***
mass         0.077860   0.020120   3.870 0.000109 ***
age          0.054075   0.013236   4.085 4.40e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 498.10  on 391  degrees of freedom
Residual deviance: 354.37  on 388  degrees of freedom
AIC: 362.37

Number of Fisher Scoring iterations: 5
```

© Chair of Decision Sciences and Systems,
Technical University of Munich

DSS
Decision Sciences & Systems

Tutorial Business Analytics

Agenda

1. Generalized Linear Models
2. Logistic Regression
3. Poisson Regression
4. Maximum Likelihood Estimation
5. Evaluation and Goodness-of-Fit

© Chair of Decision Sciences and Systems,
Technical University of Munich