

# Regressions- och tidsserieanalys

## Föreläsning 11 - Multipel logistisk regression.

**Mattias Villani** 🧑

Statistiska institutionen  
Stockholms universitet

Institutionen för datavetenskap  
Linköpings universitet



- Multipel logistisk regression
- Estimation av logistisk regression

# Multipel logistisk regression

- **Multipel logistisk regression** med  $k$  förklarande variabler:

$$P(y = 1|x_1, \dots, x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)}$$

- **Odds**

$$\text{Odds}(y = 1|\text{alla } x) = \frac{P(y = 1|\text{alla } x)}{P(y = 0|\text{alla } x)} = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k)$$

- **Oddsquot**

$$\text{OR}_j = \frac{\text{Odds}(y = 1|x_j + 1, \text{allt annat lika})}{\text{Odds}(y = 1|x_j, \text{allt annat lika})} = \exp(\beta_j)$$

# Vilka överlevde Titanic? Multipel logistisk

- $n = 891$  personer på Titanic, varav 342 överlevande.
- Responsvariabel:  $y = 1$  om överlevde, annars  $y = 0$ .
- Förklarande variabler:
  - ▶ Age
  - ▶ Sex (1=Kvinna, 0 = Man)
  - ▶ FirstClass (1=Första klass, 0 = Ej första klass)

```
> library(regkurs)
> fit <- glm(survived ~ age + sex + firstclass, data = titanic, family = binomial)
> logisticregsummary(fit, conf_intervals = T)
```

## Parameter estimates

```
-----
              Estimate Std. Error z value Pr(>|z|)    2.5 %    97.5 %
(Intercept) -1.190302   0.2168513 -5.4890 4.0416e-08 -1.621056 -0.769945
age          -0.027371   0.0068935 -3.9705 7.1724e-05 -0.041084 -0.014028
sexfemale     2.586333   0.1815696 14.2443 4.8637e-46  2.236692  2.949135
firstclassTRUE 1.958678   0.2274770  8.6104 7.2779e-18  1.520081  2.412879
```

## Odds ratio estimates

```
-----
              Estimate Std. Error z value Pr(>|z|)    2.5 %    97.5 %
(Intercept)  0.30413    1.2422 -5.4890 4.0416e-08 0.19769 0.46304
age           0.97300     1.0069 -3.9705 7.1724e-05 0.95975 0.98607
sexfemale    13.28098     1.1991 14.2443 4.8637e-46 9.36231 19.08943
firstclassTRUE 7.08995     1.2554  8.6104 7.2779e-18 4.57259 11.16607
```

# Vilka överlevde Titanic?

Parameter estimates

	Estimate
(Intercept)	-1.190302
age	-0.027371
sexfemale	2.586333
firstclassTRUE	1.958678

Odds ratio estimates

	Estimate
(Intercept)	0.30413
age	0.97300
sexfemale	13.28098
firstclassTRUE	7.08995

## ■ Logistisk regression - Odds version

$$\text{Odds}(y = 1|x) = \exp(\beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{Sex} + \beta_3 \cdot \text{FirstClass})$$

## ■ Interceptet $\beta_0$ - Oddset överleva, nyfödd pojke, ej första klass:

$$\text{Odds}(y = 1|\text{Age} = 0, \text{Sex} = 0, \text{First} = 0) = \exp(\beta_0) = \exp(-1.190302) = 0.30413$$

$$P(y = 1|\text{Age} = 0, \text{Sex} = 0, \text{First} = 0) = \frac{\text{Odds}}{1 + \text{Odds}} = \frac{0.30413}{1 + 0.30413} = 0.23321$$

## ■ Nyfödd flicka, ej i första klass:

$$\begin{aligned}\text{Odds}(y = 1|\text{Age} = 0, \text{Sex} = 1, \text{First} = 0) &= \exp(\beta_0 + \beta_2) = \exp(\beta_0) \exp(\beta_2) \\ &= 0.30413 \cdot 13.28098 = 4.039144\end{aligned}$$

$$P(y = 1|\text{Age} = 0, \text{Sex} = 1, \text{First} = 0) = \frac{4.039144}{1 + 4.039144} = 0.8015536$$

# Vilka överlevde Titanic?

## Parameter estimates

	Estimate
(Intercept)	-1.190302
age	-0.027371
sexfemale	2.586333
firstclassTRUE	1.958678

## Odds ratio estimates

	Estimate
(Intercept)	0.30413
age	0.97300
sexfemale	13.28098
firstclassTRUE	7.08995

### ■ Nyfödd flicka, första klass:

$$\begin{aligned}\text{Odds}(y = 1 | \text{Age} = 0, \text{Sex} = 1, \text{FirstClass} = 1) &= \exp(\beta_0 + \beta_2 + \beta_3) \\ &= \exp(\beta_0) \exp(\beta_2) \exp(\beta_3) = 4.039144 \cdot 7.08995 = 28.63733\end{aligned}$$

$$P(y = 1 | \text{Age} = 0, \text{Sex} = 1, \text{FirstClass} = 1) = \frac{28.63733}{1 + 28.63733} = 0.9662588$$

### ■ 1-årig flicka, första klass:

$$\begin{aligned}\text{Odds}(y = 1 | \text{Age} = 1, \text{Sex} = 1, \text{FirstClass} = 1) &= \exp(\beta_0 + \beta_1 \cdot 1 + \beta_2 + \beta_3) \\ &= \exp(\beta_0) \exp(\beta_1) \exp(\beta_2) \exp(\beta_3) = 28.63733 \cdot 0.973 = 27.86412\end{aligned}$$

$$P(y = 1 | \text{Age} = 1, \text{Sex} = 1, \text{FirstClass} = 1) = \frac{27.86412}{1 + 27.86412} = 0.9653549$$

# Vilka överlevde Titanic?

## Parameter estimates

	Estimate
(Intercept)	-1.190302
age	-0.027371
sexfemale	2.586333
firstclassTRUE	1.958678

## Odds ratio estimates

	Estimate
(Intercept)	0.30413
age	0.97300
sexfemale	13.28098
firstclassTRUE	7.08995

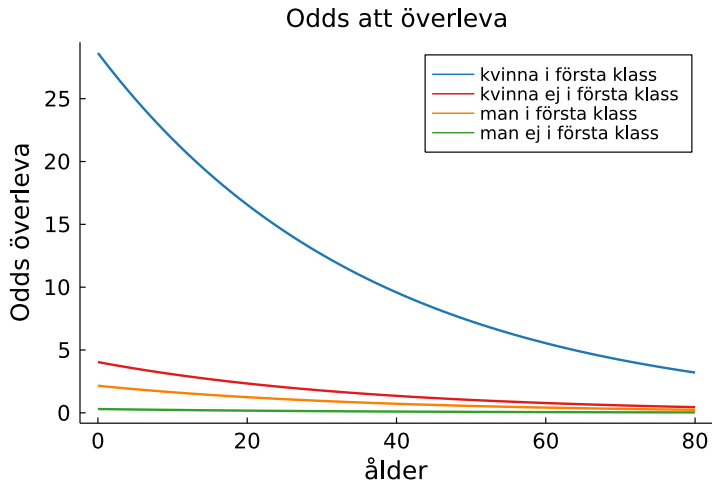
### ■ 2-årig flicka, första klass:

$$\begin{aligned}\text{Odds}(y = 1 | \text{Age} = 2, \text{Sex} = 1, \text{FirstClass} = 1) &= \exp(\beta_0 + \beta_1 \cdot 2 + \beta_2 + \beta_3) \\ &= \exp(\beta_0 + \beta_1 + \beta_1 + \beta_2 + \beta_3) \\ &= \exp(\beta_0) \exp(\beta_1) \exp(\beta_1) \exp(\beta_2) \exp(\beta_3) = 27.86412 \cdot 0.973 = 27.11179 \\ P(y = 1 | \text{Age} = 2, \text{Sex} = 1, \text{FirstClass} = 1) &= \frac{27.11179}{1 + 27.11179} = 0.9644277\end{aligned}$$

### ■ Kontroll:

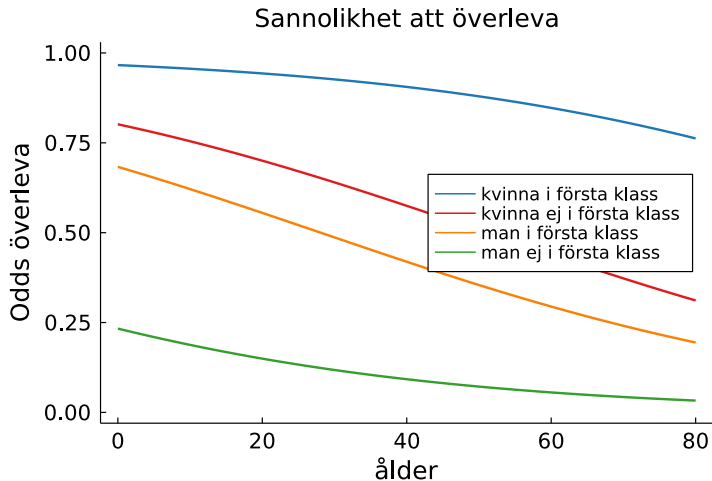
$$\begin{aligned}P(y = 1 | \text{Age} = 2, \text{Sex} = 1, \text{FirstClass} = 1) \\ &= \frac{\exp(-1.190302 - 0.027371 \cdot 2 + 2.586333 + 1.958678)}{1 + \exp(-1.190302 - 0.027371 \cdot 2 + 2.586333 + 1.958678)} = 0.9644277\end{aligned}$$

# Vilka överlevde Titanic? Odds





# Vilka överlevde Titanic? Sannolikhet



# Skatta en logistisk regression

- Datamaterial med tre **oberoende** datapunkter ( $n = 3$ ):

$$y_1 = 0, y_2 = 1, y_3 = 0.$$

- Varje  $y_i$  observeras tillsammans med en förklarande variabel

$$x_1, x_2, x_3$$

- **Sannolikheten för just detta datamaterial:**

$$\underbrace{\frac{1}{1 + \exp(\beta_0 + \beta_1 x_1)}}_{y_1=0} \cdot \underbrace{\frac{\exp(\beta_0 + \beta_1 x_2)}{1 + \exp(\beta_0 + \beta_1 x_2)}}_{y_2=1} \cdot \underbrace{\frac{1}{1 + \exp(\beta_0 + \beta_1 x_3)}}_{y_3=0}$$

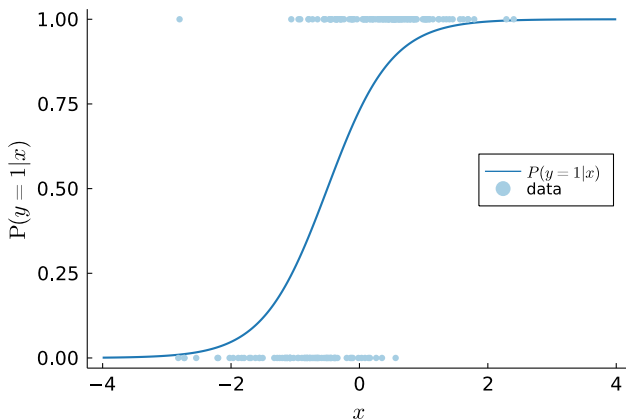
- **Likelihoodfunktion:** sannolikheten för ett datamaterial betraktat som en funktion av parametrarna  $\beta_0$  och  $\beta_1$ .
- **Maximum likelihood:** välj de parametervärden  $\beta_0$  och  $\beta_1$  som maximerar sannolikheten för det observerade datamaterialet.

# Logistisk regression - maximum likelihood

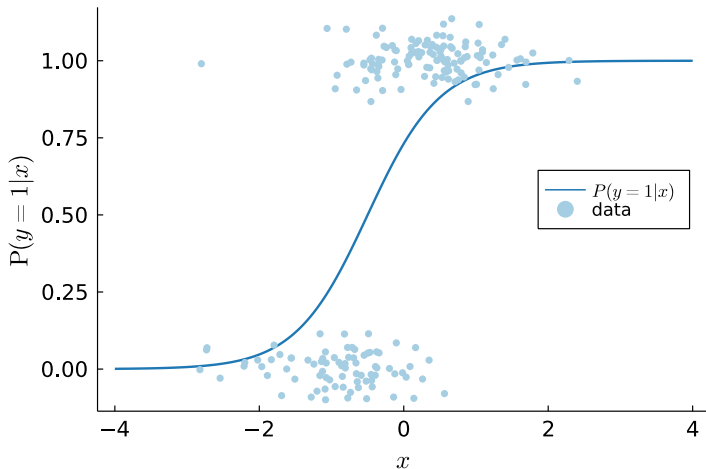
■ Data  $(x_1, y_1), \dots, (x_n, y_n)$  simulerat från logistisk regression.

▶  $\beta_0 = 1$  och  $\beta_1 = 2$

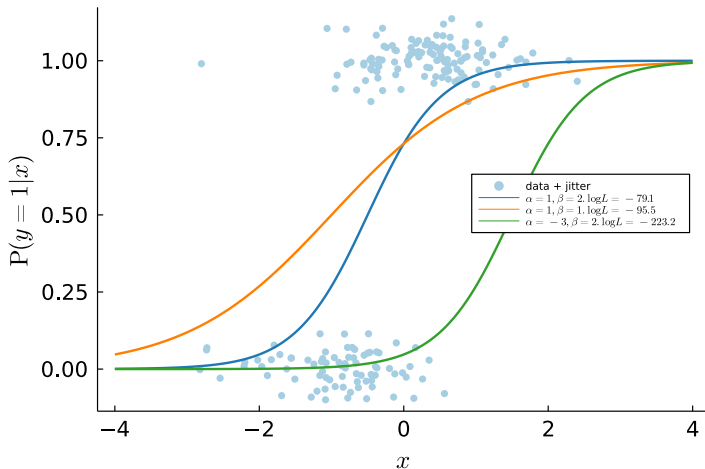
▶  $n = 200$



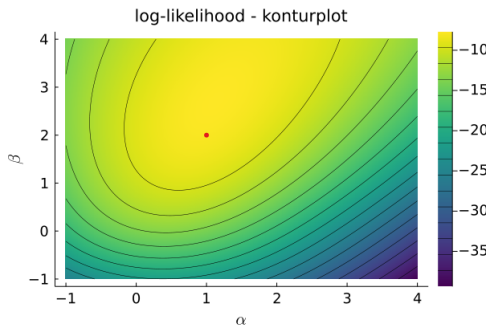
# Skatta en logistisk regression - jitter



# Skatta en logistisk regression - jitter



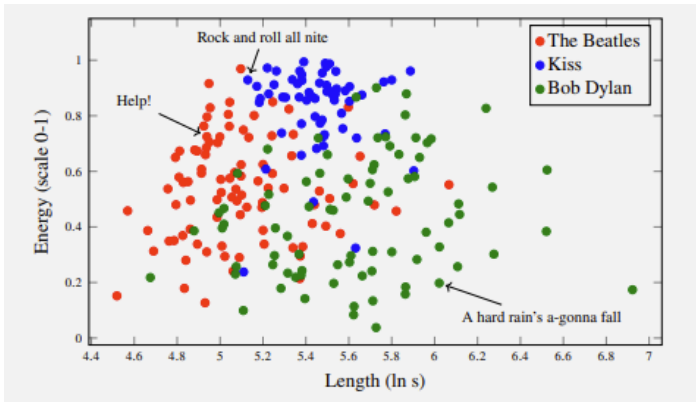
# Skatta en logistisk regression



# Multinomial logistisk regression

## ■ Spotify-data från boken

Machine Learning - A First Course for Engineers and Scientists



## ■ Respons med fler än två kategorier (binärt). **Multinomial logistisk regression.**