Regressions- och tidsserieanalys

Föreläsning 11 - Logistisk regression.

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet

mattiasvillani.com @matvil mattiasvillani

Översikt

- Odds och logodds
- Enkel logistisk regression
- Multipel logistisk regression
- Estimation av logistisk regression

Odds och logodds

- Låt $\mathrm{P}(A)$ vara **sannolikheten** för en händelse A.

$$\mathrm{P}(A) = \frac{\text{antal fall där } A \text{ inträffar}}{\text{antal möjliga fall}}$$

- **Odds**

$$\mathrm{Odds}(A) = \frac{\text{antal fall där } A \text{ inträffar}}{\text{antal fall där A inte inträffar}}$$

$$\mathrm{Odds}(A) = \frac{\mathrm{P}(A)}{1 - \mathrm{P}(A)}$$

- Exempel: Sannolikheten att slå en 6:a med en vanlig tärning:
  - Sannolikhet $\mathrm{P}(A) = 1/6$
  - Odds

$$\mathrm{Odds}(A) = \frac{1/6}{5/6} = \frac{1}{5}$$

Oddset är $1:5$ (“1mot 5”).

Exponentialfunktionen

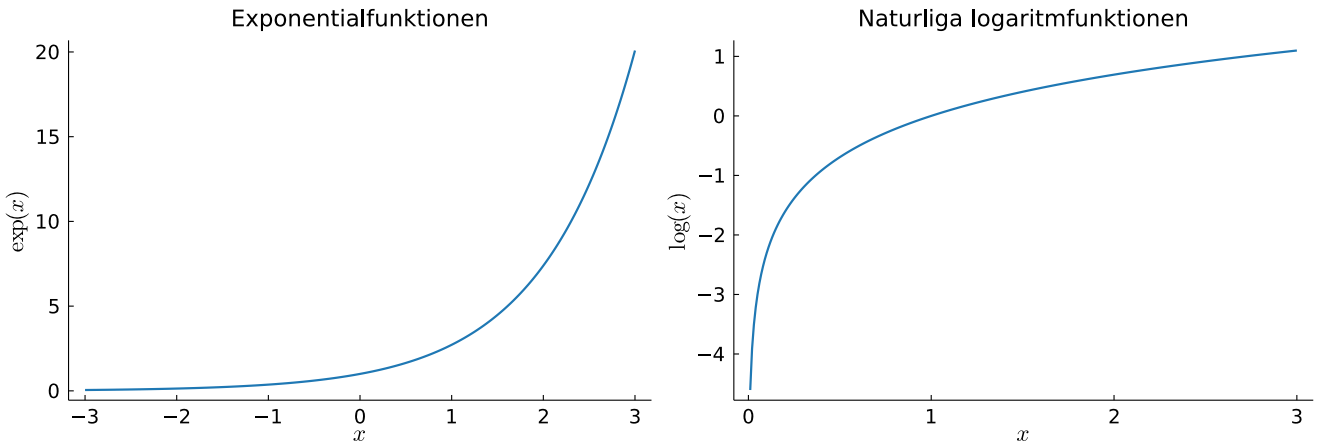
Exponentialfunktionen

$$\exp(x) = e^x$$

där $e \approx 2.71828$ är **Eulers tal**.

- **Naturliga logaritmen** $\ln(x)$ är inversa funktionen till $\exp(x)$.

$$\ln(\exp(x)) = \ln(e^x) = x$$

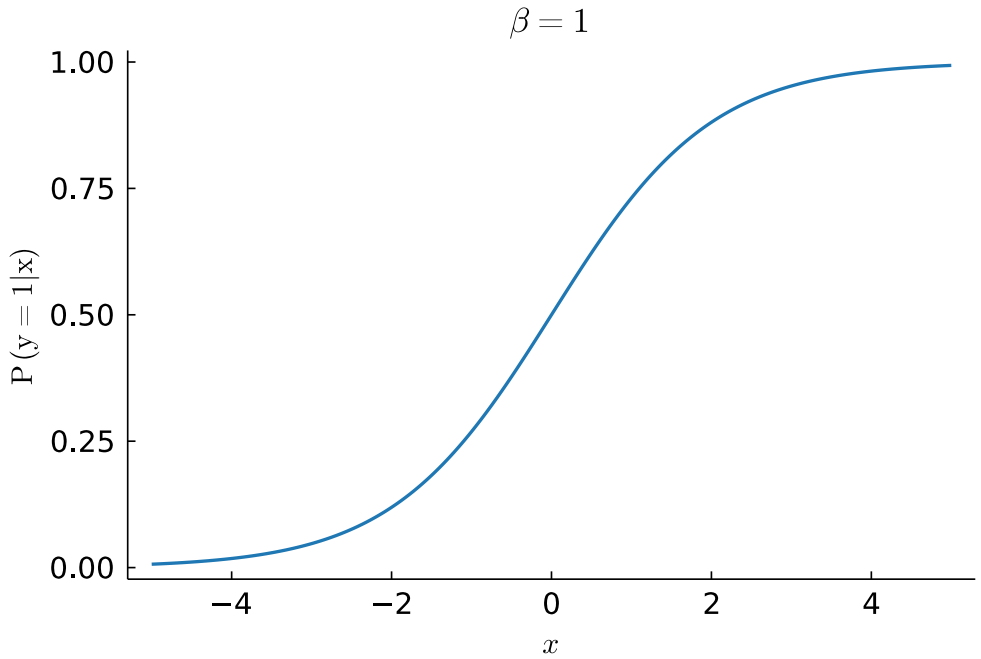


Logistisk regression - sannolikhet för $y = 1$

- Binär responsvariabel: $y = 0$ och $y = 1$.
- Logistisk regression

$$\mathrm{P}(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$\mathrm{P}(y = 0|x) = 1 - \mathrm{P}(y = 1|x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$



Logistisk regression - oddskvot

- Logistisk regression

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$P(y = 0|x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

- Odds

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

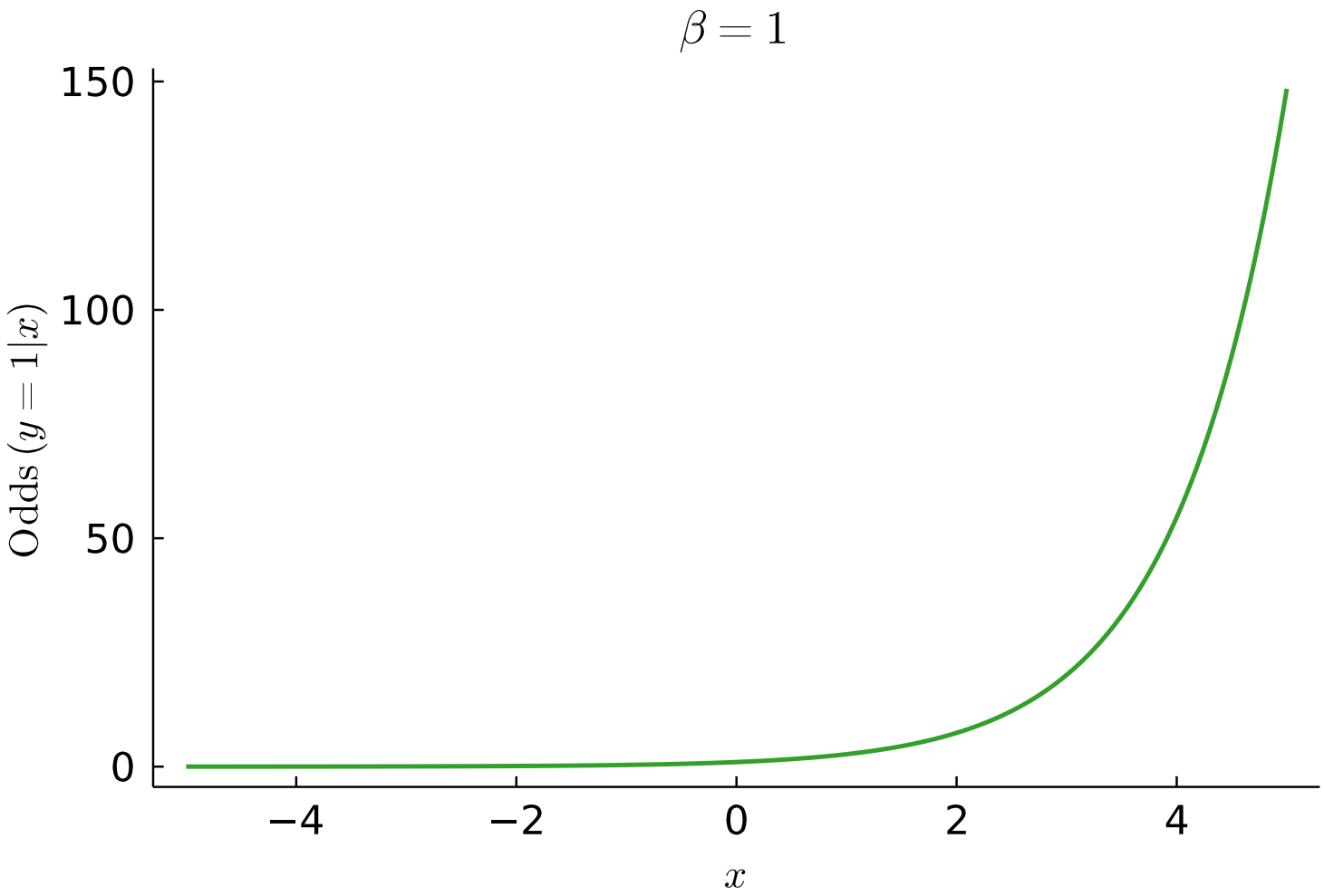
- **Oddskvot** för att tolka β_1

$$\text{OR}(x) = \frac{\text{Odds}(y = 1|x + 1)}{\text{Odds}(y = 1|x)} = \exp(\beta_1)$$

- Bevis:

$$\text{OR}(x) = \frac{\text{Odds}(y = \mathbf{1}|x + \mathbf{1})}{\text{Odds}(y = \mathbf{1}|x)} = \frac{\exp(\beta_\mathbf{0} + \beta_\mathbf{1} x + \beta_\mathbf{1})}{\exp(\beta_\mathbf{0} + \beta_\mathbf{1} x)} = \frac{\exp(\beta_\mathbf{0} + \beta_\mathbf{1} x)\exp(\beta_\mathbf{1})}{\exp(\beta_\mathbf{0} + \beta_\mathbf{1} x)} = \exp(\beta_\mathbf{1})$$

Logistisk regression - oddskvot



Logistisk regression - log-odds

- Repetition: Logaritm med bas 10:

$$\log(10^a) = a$$

- **Naturlig logaritm** (bas $e \approx 2.7183$)

$$\ln(\exp(a)) = \ln e^a = a$$

- Logistisk regression

$$\mathrm{P}(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- Odds

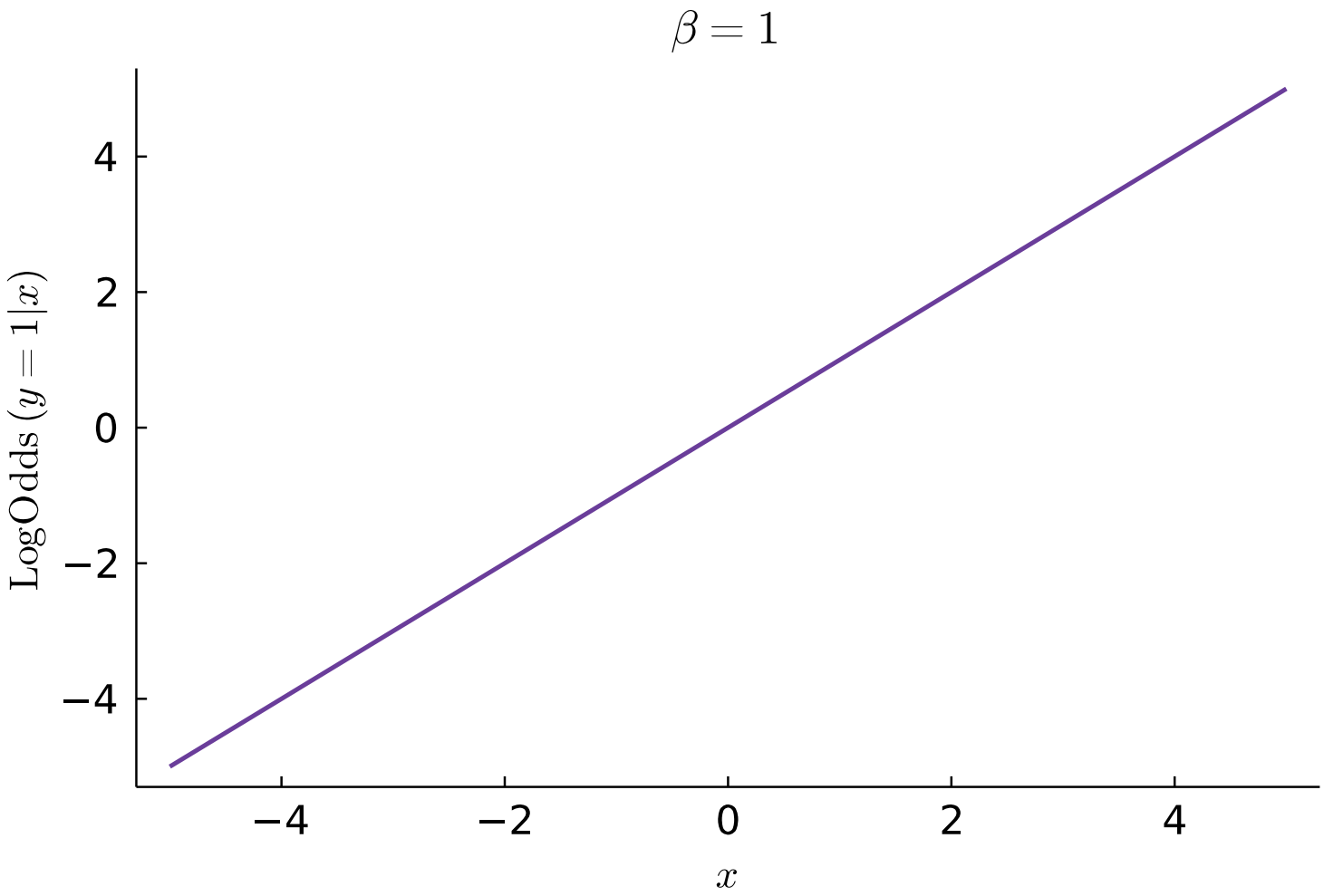
$$\mathrm{Odds}(y = 1|x) = \exp(\beta_0 + \beta_1 x)$$

- **Log-odds**

$$\mathrm{LogOdds}(y = 1|x) = \beta_0 + \beta_1 x$$

- Logistisk regression är en **linjär modell** för log-oddset.

Logistisk regression - logodds



Enkel logistisk regression - Wisconsin Cancer

- $n = 699$ samples klassificerade som
  - 'benign' (458 fall)
  - 'malignant' (241 fall).
- Ett antal mått (ordinalskala 1-10) som förklarande variabler:

Variabelnamn	Mått
Cl.thickness	Clump Thickness
Cell.size	Uniformity of Cell Size
Cell.shape	Uniformity of Cell Shape
Marg.adhesion	Marginal Adhesion
Epith.c.size	Single Epithelial Cell Size
Bl.cromatin	Bland Chromatin
Normal.nucleoli	Normal Nucleoli
Mitoses	Mitoses

	Id	Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size	Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1	1000025	5	1	1	1	2	1	3	1	1	benign
2	1002945	5	4	4	5	7	10	3	2	1	benign
3	1015425	3	1	1	1	2	2	3	1	1	benign
4	1016277	6	8	8	1	3	4	3	7	1	benign
5	1017023	4	1	1	3	2	1	3	1	1	benign
6	1017122	8	10	10	8	7	10	9	7	1	malignant

Enkel logistisk regression - Wisconsin Cancer

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.9602	0.3600	189.8772	<.0001
Cellsize	1	1.4887	0.1210	151.2705	<.0001

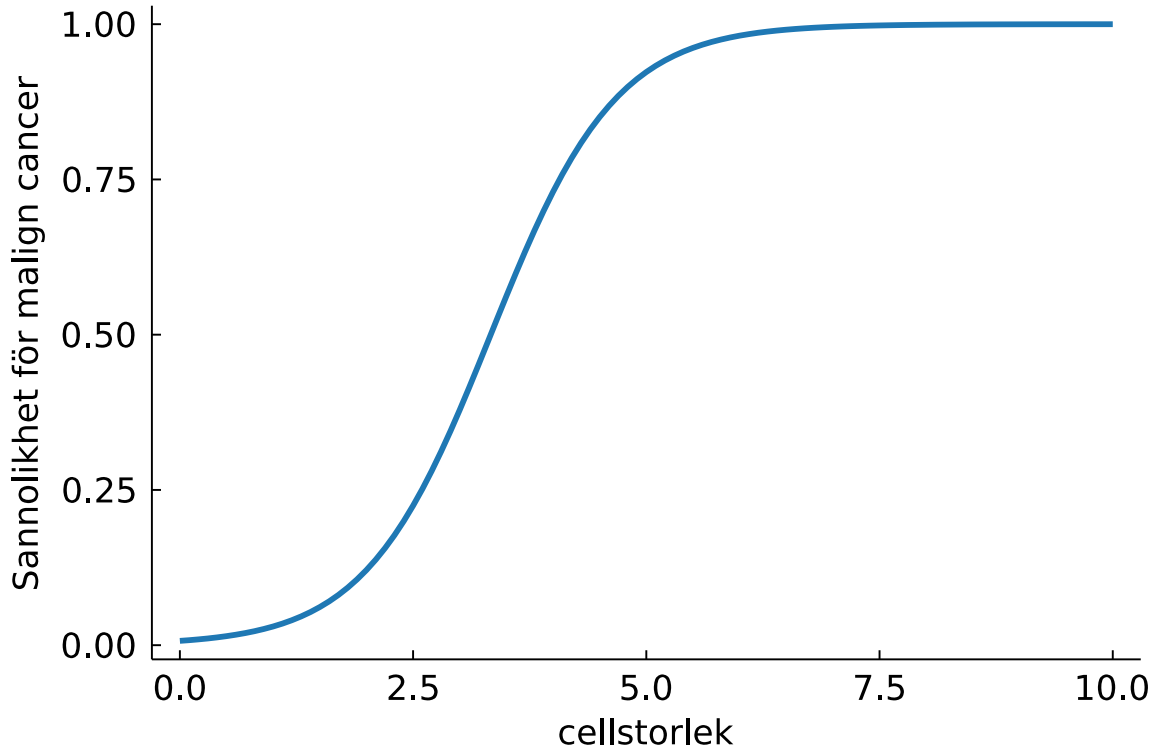
Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Cellsize	4.431	3.495	5.618

```
title "Simple logistic regression";
proc logistic data = work.breastcancer DESCENDING;
model class = Cellsize;
run;
```

- Ökning av `Cellsize` med en enhet ökar risken (oddset) för malign $\exp(1.4887) = 4.431$ ggr, eller med 443.1%.

Enkel logistisk regression - Wisconsin Cancer



Multipel logistisk regression

- Multipel logistisk regression med k förklarande variabler:

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

- Logodds

$$\text{LogOdds}(y = 1|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- Oddskvot - multipel regression

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

Multipel logistisk regression - Wisconsin Cancer

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-10.1099	1.1737	74.1902	<.0001
Clthickness	1	0.5352	0.1419	14.2204	0.0002
Cellsize	1	-0.00594	0.2092	0.0008	0.9773
Cellshape	1	0.3221	0.2306	1.9507	0.1625
Margadhesion	1	0.3307	0.1235	7.1742	0.0074
Epithcsize	1	0.0968	0.1566	0.3822	0.5364
Barenuclei	1	0.3830	0.0939	16.6503	<.0001
Blcromatin	1	0.4474	0.1714	6.8140	0.0090
Normalnucleoli	1	0.2131	0.1129	3.5622	0.0591
Mitoses	1	0.5385	0.3256	2.7352	0.0982

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Clthickness	1.708	1.293	2.256
Cellsize	0.994	0.660	1.498
Cellshape	1.380	0.878	2.169
Margadhesion	1.392	1.093	1.773
Epithcsize	1.102	0.811	1.497
Barenuclei	1.467	1.220	1.763
Blcromatin	1.564	1.118	2.189
Normalnucleoli	1.237	0.992	1.544
Mitoses	1.713	0.905	3.244

- `Cellsize` inte signifikant (oddskvotens KI innehåller värdet 1).
- Om `Clthinkness` ökar med en enhet så ökar oddskvoten för `malign` med 1.708, dvs risken för `malign` ökar med 70.8%.

Modellutvärdering - Wisconsin Cancer

- **Hur ofta predikterar** den skattade modellen **rätt?**
- Enkel logistisk regression - cellsize:

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	96.1	Somers' D	0.948
Percent Discordant	1.3	Gamma	0.973
Percent Tied	2.6	Tau-a	0.429
Pairs	110378	c	0.974

- Multipel logistisk regression

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	99.6	Somers' D	0.993
Percent Discordant	0.4	Gamma	0.993
Percent Tied	0.0	Tau-a	0.452
Pairs	106116	c	0.996

- Bättre att göra detta på en **träning-test split** av data.

Vilka överlevde Titanic?

- $n = 891$ 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)

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1499	0.2291	25.2034	<.0001
Age	1	-0.0283	0.00723	15.3758	<.0001
Sex	1	2.6052	0.1815	206.0644	<.0001
FirstClass	1	1.8960	0.2197	74.4654	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.972	0.958	0.986
Sex	13.534	9.483	19.316
FirstClass	6.660	4.329	10.244

Vilka överlevde Titanic?

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.972	0.958	0.986
Sex	13.534	9.483	19.316
FirstClass	6.660	4.329	10.244

- Logistisk regression - Odds version

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

- Interceptet β_0 - Oddset överleva, nyfödd pojke, ej första klass:

$$\text{Odds}(y = 1|\texttt{Age} = 0, \texttt{Sex} = 0, \texttt{First} = 0) = \exp(\beta_0) = \exp(-1.1499) = 0.3166684$$

$$\text{P}(y = 1|\texttt{Age} = 0, \texttt{Sex} = 0, \texttt{First} = 0) = \frac{\text{Odds}}{1 + \text{Odds}} = \frac{0.3166684}{1 + 0.3166684} = 0.2405073$$

- Nyfödd flicka, ej i första klass:

$$\begin{aligned}\text{Odds}(y = 1|\texttt{Age} = 0, \texttt{Sex} = 1, \texttt{First} = 0) &= \exp(\beta_0 + \beta_2) = \exp(\beta_0)\exp(\beta_2) \\ &= 0.3166684 \cdot 13.543 = 4.28864\end{aligned}$$

$$\text{P}(y = 1|\texttt{Age} = 0, \texttt{Sex} = 1, \texttt{First} = 0) = \frac{4.28864}{1 + 4.28864} = 0.8109155$$

Vilka överlevde Titanic?

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.972	0.958	0.986
Sex	13.534	9.483	19.316
FirstClass	6.660	4.329	10.244

- Nyfödd flicka, första klass:

$$\begin{aligned}\text{Odds}(y = 1|\texttt{Age} = 0, \texttt{Sex} = 1, \texttt{FirstClass} = 1) &= \exp(\beta_0 + \beta_2 + \beta_3) \\ &= \exp(\beta_0)\exp(\beta_2)\exp(\beta_3) = 4.28864 \cdot 6.66 = 28.56234\end{aligned}$$

$$\text{P}(y = 1|\texttt{Age} = 0, \texttt{Sex} = 1, \texttt{FirstClass} = 1) = \frac{28.56234}{1 + 28.56234} = 0.9661732$$

- 1-årig flicka, första klass:

$$\begin{aligned}\text{Odds}(y = 1|\texttt{Age} = 1, \texttt{Sex} = 1, \texttt{FirstClass} = 1) &= \exp(\beta_0 + \beta_1 \cdot 1 + \beta_2 + \beta_3) \\ &= \exp(\beta_0)\exp(\beta_1)\exp(\beta_2)\exp(\beta_3) = 28.56234 \cdot 0.972 = 27.7626\end{aligned}$$

$$\text{P}(y = 1|\texttt{Age} = 1, \texttt{Sex} = 1, \texttt{FirstClass} = 1) = \frac{27.76264}{1 + 27.7626} = 0.9652326$$

Vilka överlevde Titanic?

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.972	0.958	0.986
Sex	13.534	9.483	19.316
FirstClass	6.660	4.329	10.244

- 2-årig flicka, första klass:

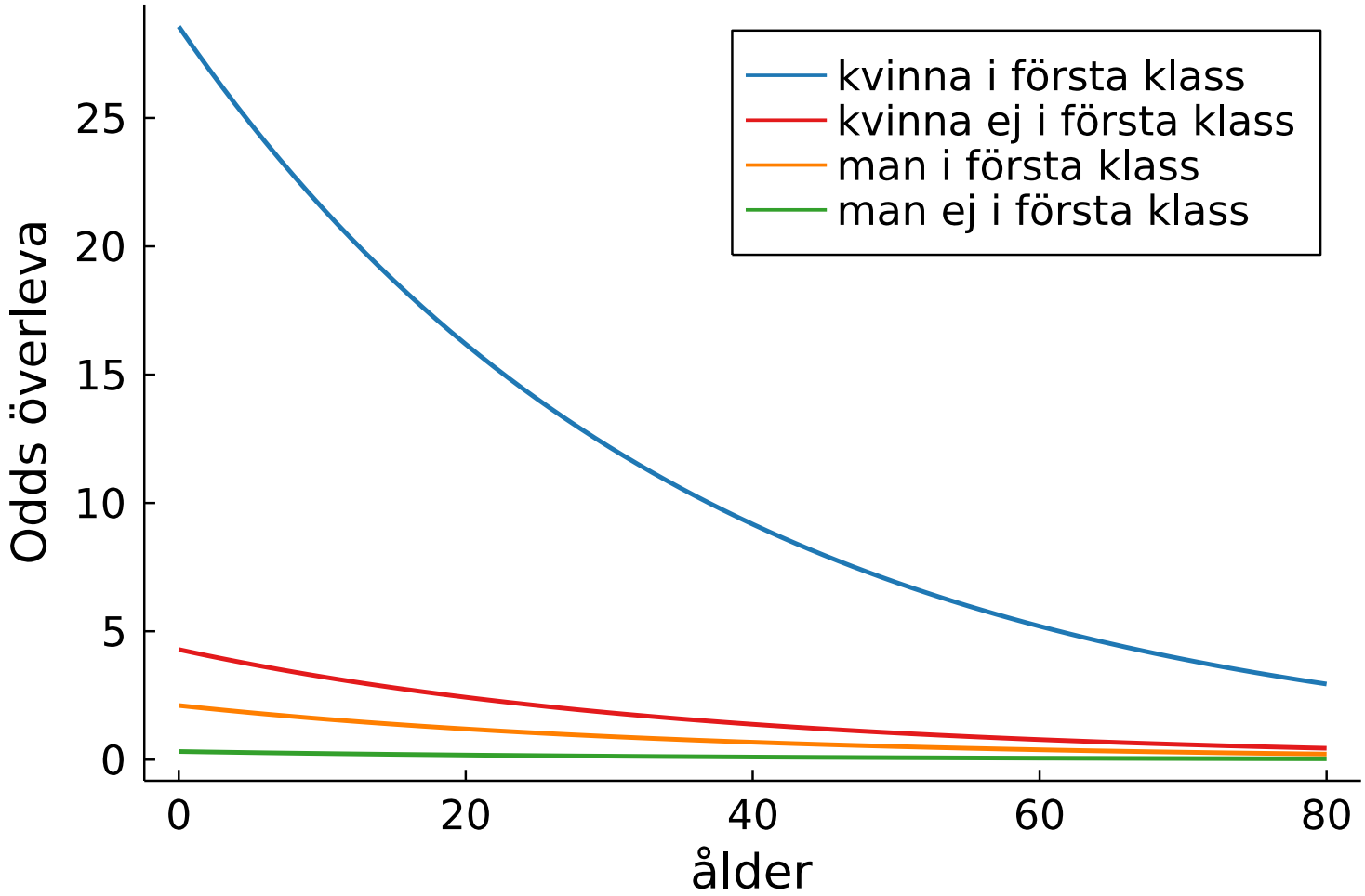
$$\begin{aligned}\text{Odds}(y = 1|\texttt{Age} = 2, \texttt{Sex} = 1, \texttt{FirstClass} = 1) &= \exp(\beta_0 + \beta_1 \cdot 2 + \beta_2 + \beta_3) \\ &= \exp(\beta_0 + \beta_1 + \beta_1 + \beta_2 + \beta_3) \\ &= \exp(\beta_0)\exp(\beta_1)\exp(\beta_1)\exp(\beta_2)\exp(\beta_3) = 27.7626 \cdot 0.972 = 26.98525\end{aligned}$$

$$\text{P}(y = 1|\texttt{Age} = 2, \texttt{Sex} = 1, \texttt{FirstClass} = 1) = \frac{26.98525}{1 + 26.98525} = 0.9642669$$

- Kontroll:

$$\begin{aligned}&\text{P}(y = 1|\texttt{Age} = 2, \texttt{Sex} = 1, \texttt{FirstClass} = 1) \\ &= \frac{\exp(-1.1499 - 0.0283 \cdot 2 + 2.6052 + 1.896)}{1 + \exp(1 - 1.1499 - 0.0283 \cdot 2 + 2.6052 + 1.896)} = 0.9642465\end{aligned}$$

Vilka överlevde Titanic?



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

$$x_1, x_2, x_3$$

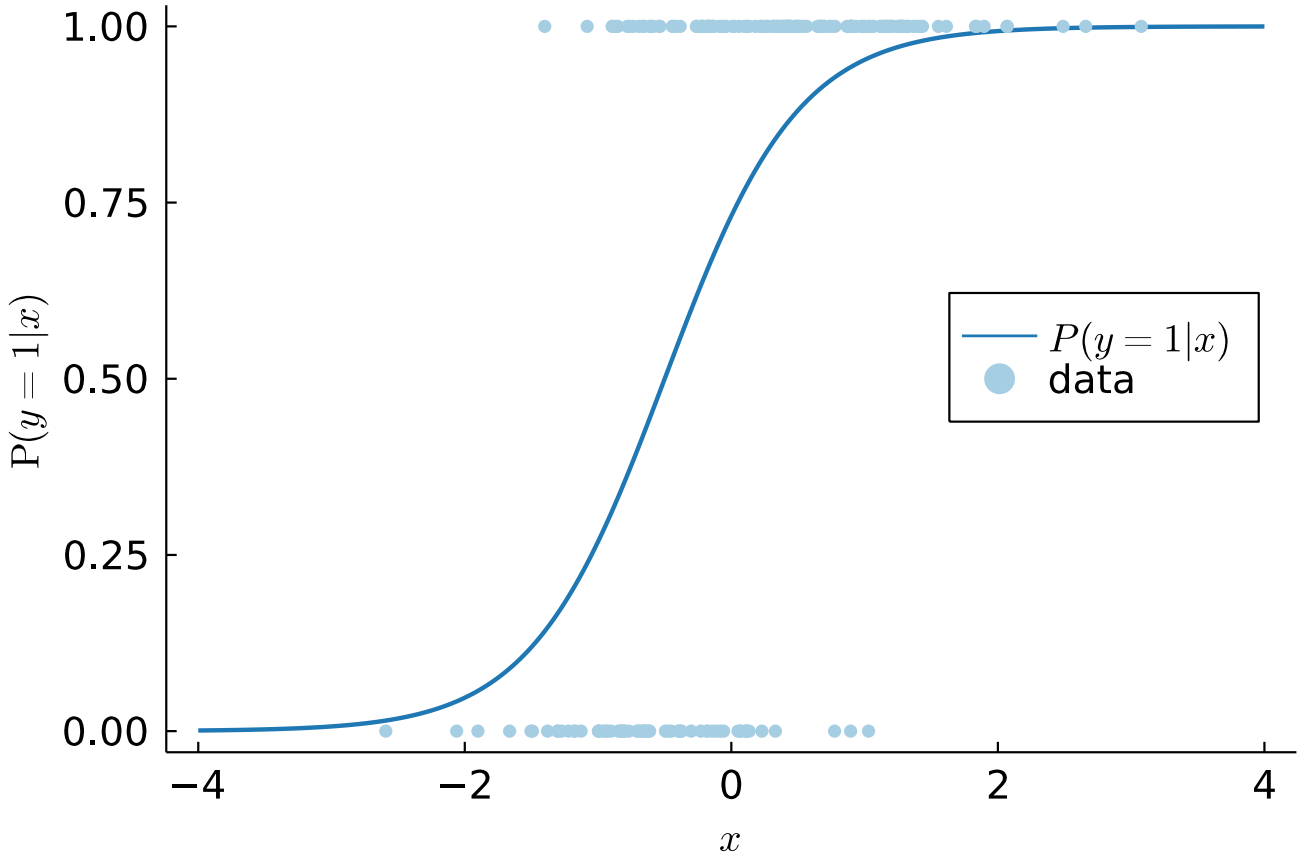
- **Sannolikheten för just detta datamaterial**:

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

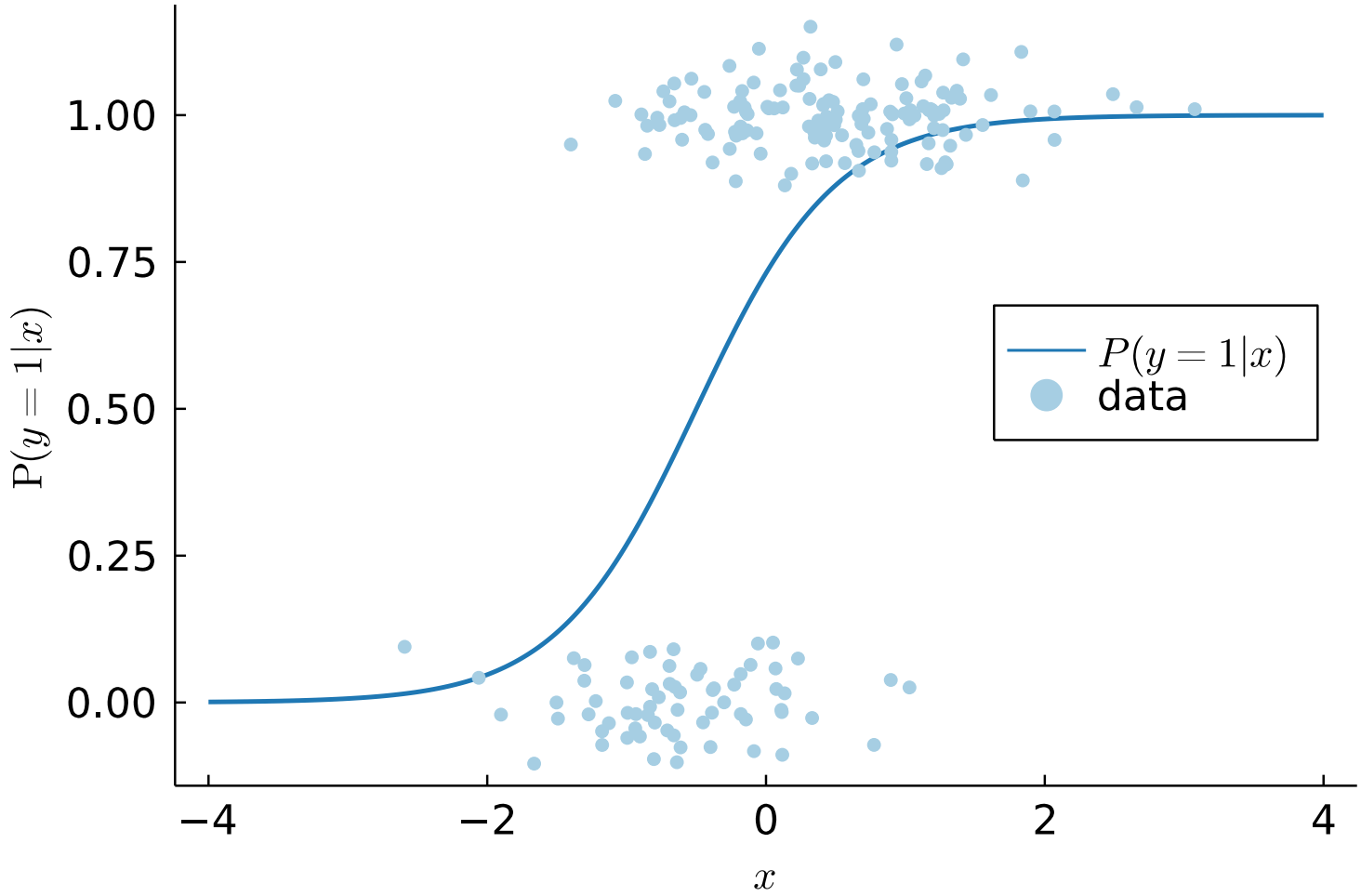
- **Likelihoodfunktion**: sannolikheten för ett datamaterial betraktat som en funktion av parametrarna β_0 och β_1.
- **Maximum likelihood**: välj de parametervärden β_0 och β_1 som maximerar sannolikheten för det observerade datamaterialet.

Logistisk regression - maximum likelihood

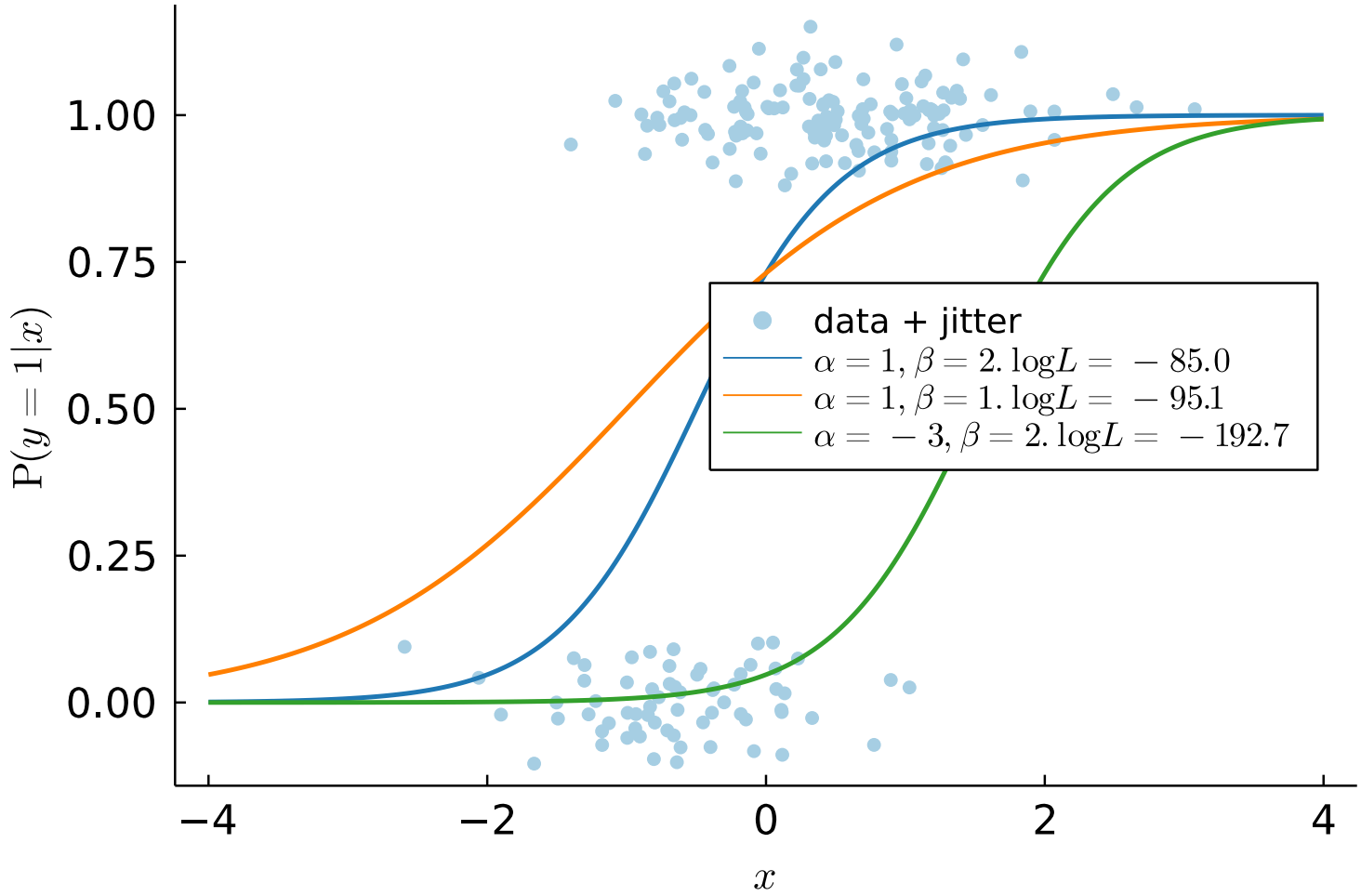
- Data $(x_1, y_1), \ldots, (x_n, y_n)$ simulerat från logistisk regression.
  - $\alpha = 1$ och $\beta = 2$
  - $n = 200$



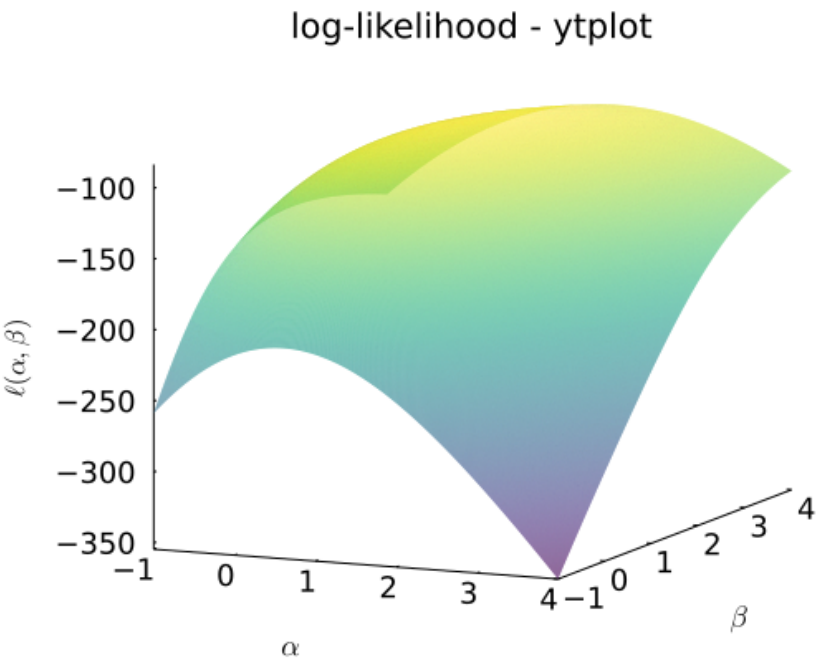
Skatta en logistisk regression - jitter

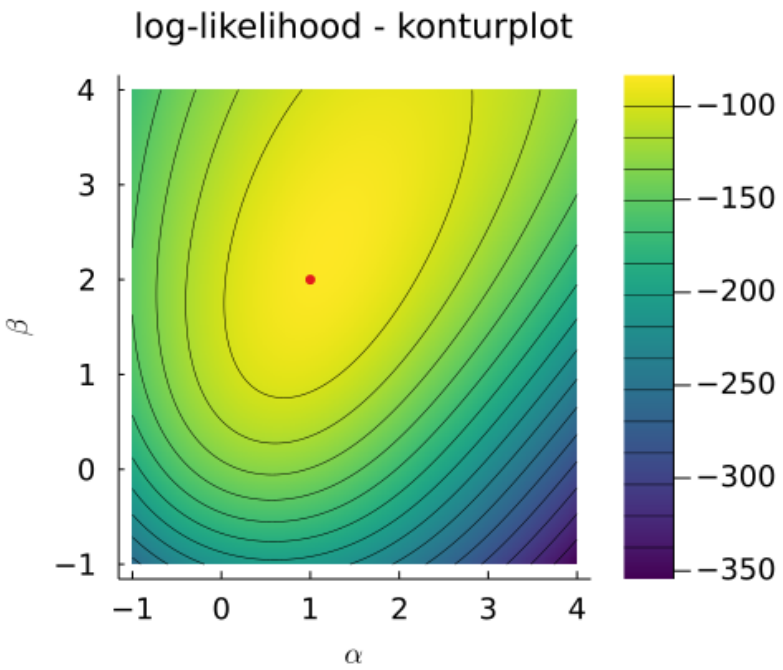


Skatta en logistisk regression - jitter



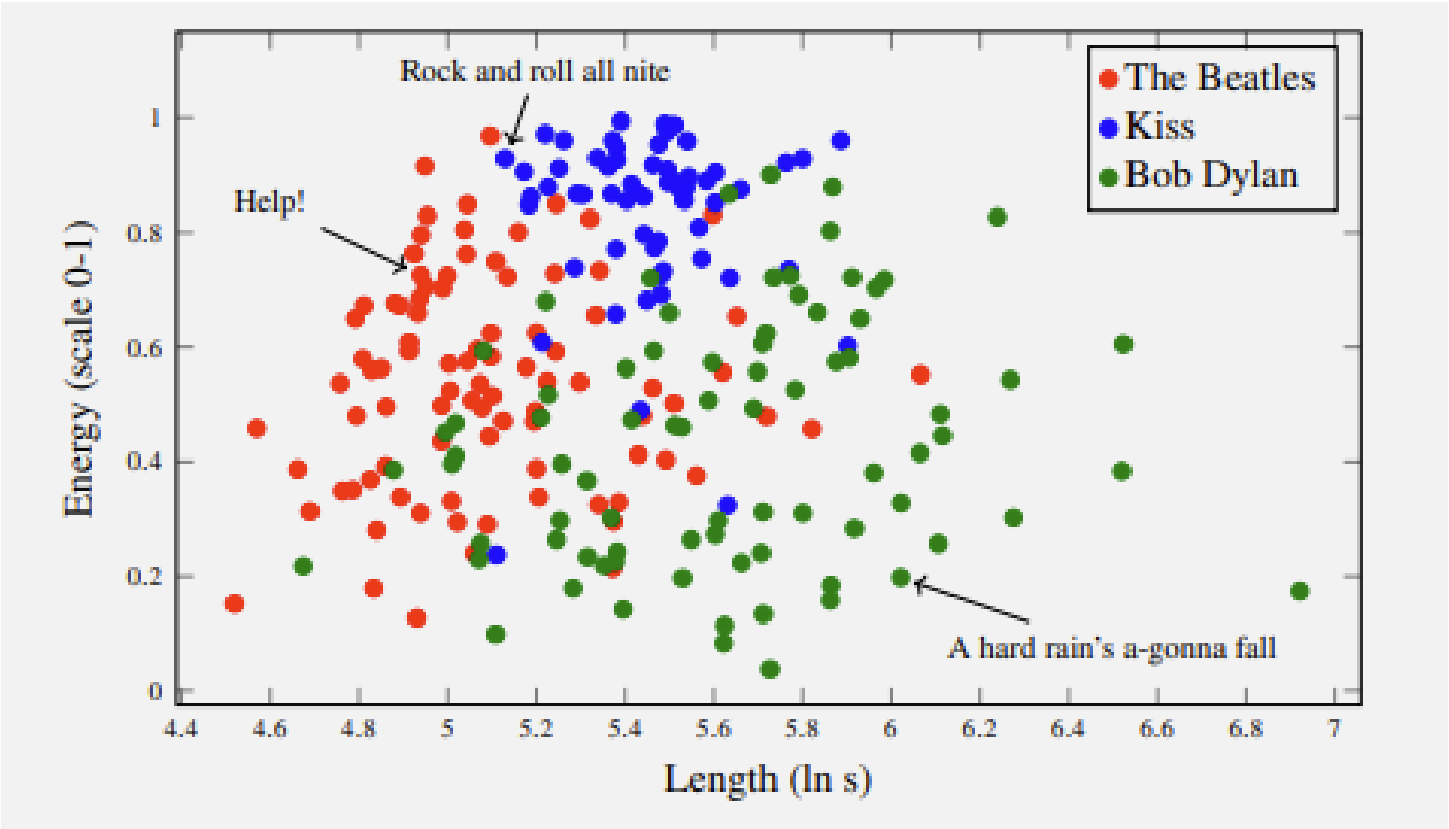
Skatta en logistisk regression





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**.

Klassificeringsyta - spotify data