

# Facit laboration 3

Mona Sfaxi

## Del 1 - Skatta en logistisk regressionsmodell

### 1a - Testdata och träningsdata

Vi väljer ut testdata. Jag börjar med att sätta ett seed till 4910 så att jag kan replikera resultatet, så att samma observationer väjs ut till test- och träningsdata.

**Exempel på testdata givet att man har använt ett så kallat seed på 4910**

	name	survived	sex	age
115	Mr. Edvard Pekoniemi	0	male	21
784	Master. George Hugh Rice	0	male	8
211	Miss. Clear Annie Cameron	1	female	35
201	Mr. Frederick Sage	0	male	17
9	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson	1	female	27
418	Miss. Catharina Van Impe	0	female	10

  

	fare	firstclass
115	7.9250	FALSE
784	29.1250	FALSE
211	21.0000	FALSE
201	69.5500	FALSE
9	11.1333	FALSE
418	24.1500	FALSE

(Lägg märke till de olika radnumren och att de inte består av ordnade värden från 1 till 89).

## Exempel på träningsdata givet att man har använt ett seed på 4910

	name	survived	sex	age	fare
1	Mr. Owen Harris Braund	0	male	22	7.2500
3	Miss. Laina Heikkinen	1	female	26	7.9250
4	Mrs. Jacques Heath (Lily May Peel) Futrelle	1	female	35	53.1000
5	Mr. William Henry Allen	0	male	35	8.0500
6	Mr. James Moran	0	male	27	8.4583
7	Mr. Timothy J McCarthy	0	male	54	51.8625

  

	firstclass
1	FALSE
3	FALSE
4	TRUE
5	FALSE
6	FALSE
7	TRUE

## 1b - Logistisk regression

Estimat för den mindre modellen:

```
logisticreg_summary(mod_mini)
```

Parameter estimates

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.132303	0.224943	-5.0337	4.8105e-07
sexfemale	2.508496	0.188307	13.3213	1.7409e-40
age	-0.027508	0.007165	-3.8392	1.2345e-04
firstclassTRUE	1.911255	0.237849	8.0356	9.3136e-16

Odds ratio estimates

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.32229	1.2523	-5.0337	4.8105e-07
sexfemale	12.28644	1.2072	13.3213	1.7409e-40
age	0.97287	1.0072	-3.8392	1.2345e-04
firstclassTRUE	6.76157	1.2685	8.0356	9.3136e-16

Estimat för den större modellen:

```
logisticreg_summary(mod_maxi)
```

Parameter estimates

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.04476259	0.39996935	0.11192	9.1089e-01
sexfemale	0.71388020	0.47098974	1.51570	1.2959e-01
age	-0.07044280	0.01428514	-4.93119	8.1728e-07
fare	-0.01114414	0.01046183	-1.06522	2.8678e-01
firstclassTRUE	1.48140172	0.96850772	1.52957	1.2612e-01
sexfemale:age	0.06814605	0.01719285	3.96363	7.3820e-05
sexfemale:fare	-0.00215638	0.00688953	-0.31299	7.5429e-01
sexfemale:firstclassTRUE	0.89649173	1.01423314	0.88391	3.7674e-01
age:fare	0.00025449	0.00021433	1.18737	2.3508e-01
age:firstclassTRUE	0.01639605	0.02413780	0.67927	4.9697e-01
fare:firstclassTRUE	0.00330945	0.00953759	0.34699	7.2860e-01

Odds ratio estimates

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.04578	1.4918	0.11192	9.1089e-01
sexfemale	2.04190	1.6016	1.51570	1.2959e-01
age	0.93198	1.0144	-4.93119	8.1728e-07
fare	0.98892	1.0105	-1.06522	2.8678e-01
firstclassTRUE	4.39911	2.6340	1.52957	1.2612e-01
sexfemale:age	1.07052	1.0173	3.96363	7.3820e-05
sexfemale:fare	0.99785	1.0069	-0.31299	7.5429e-01
sexfemale:firstclassTRUE	2.45099	2.7572	0.88391	3.7674e-01
age:fare	1.00025	1.0002	1.18737	2.3508e-01
age:firstclassTRUE	1.01653	1.0244	0.67927	4.9697e-01
fare:firstclassTRUE	1.00331	1.0096	0.34699	7.2860e-01

Resultatet kan skilja sig åt beroende på vilken kategori som kodas som 1 för de kategoriska variablerna. Här kodas exempelvis första klass som 1 (annars 0).

loglikelihoodvärdet för den lilla modellen ges av

```
[1] -382.2347
```

Och för de större modellen:

[1] -368.8806

Det kritiska värdet ges av:

[1] 14.06714

Vi förkastar  $H_0$  på en 5%-ig signifikansnivå ifall det observerade värdet är större än det kritiska.  
Det observerade värdet ges av:

[1] 26.70802

Så vi förkastar  $H_0$ . Testet indikerar att den större modellen är bättre.

## Del 2 - Klassifikation med logistisk regression

### 2a - Beräkna skattade sannolikheter för testdata

Prediktionerna är baserade på test- och träningsdata med ett seed = 4910

		name	survived	sex	age
115		Mr. Edvard Pekoniemi	0	male	21
784		Master. George Hugh Rice	0	male	8
211		Miss. Clear Annie Cameron	1	female	35
201		Mr. Frederick Sage	0	male	17
9	Mrs. Oscar W (Elisabeth Vilhelmina Berg) Johnson		1	female	27
418		Miss. Catharina Van Impe	0	female	10

  

	fare	firstclass	slh_mini	slh_maxi
115	7.9250	FALSE	0.1531688	0.1853520
784	29.1250	FALSE	0.2054843	0.3134464
211	21.0000	FALSE	0.6019099	0.6424462
201	69.5500	FALSE	0.1679913	0.1642499
9	11.1333	FALSE	0.6532803	0.6513667
418	24.1500	FALSE	0.7504695	0.6167858

## 2b - Förvirringsmatriser

Exempel på förvirringsmatris för den mindre modellen (resultat kan ibland skilja sig åt, även om man har satt samma seed):

```
survived  0  1
          0 51  6
          1  5 27
```

**Specificitet:**

```
[1] 89.47368
```

**Sensitivitet:**

```
[1] 84.375
```

**Förvirringsmatris för den större modellen**

```
survived  0  1
          0 51  6
          1  7 25
```

## Del 3 - Modellutvärdering baserat på prediktionsvärden

**Summan av de logaritmerade prediktionsvärdena för den lilla modellen**

```
[1] -34.363
```

**Summan av de logaritmerade prediktionsvärdena för den stora modellen**

```
[1] -34.58442
```

### Summan av de logaritmerade prediktionsvärdena för den naiva modellen

[1] -61.6901

- Utifrån resultaten ovan är den bästa modellen den lilla modellen, följt av den större och sist kommer den naiva modellen. Lägg dock märke till att detta resultat och vilken modell som bedöms som bäst utifrån prediktionsvärden kan ändras ifall man använder ett annat seed än 4910.