Labbrapport i Statistik

Laboration 1

732G54

Zerui Wang
Viet Tien Trinh

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2024-11-08

Table of contents

1 Introduktion

Detta är ett arbete som handlar om logiska regressionen, vilket innehåller totalt fem uppgifter som geomfördes av oss.

2 Labborationsuppgifter

2.1 a) Dela upp hela datamaterialet i en tränings- och valideringsmängd med 2/3 av observationerna i träningsmängden. Presentera vilket seed som använts för uppdelningen. Ta sedan fram en frekvenstabell för responsvariabeln ur båda datamängderna och kommentera på resul-tatet; hur är balansen mellan klasserna, hur stor andel finns i varje klass i de två mängderna?

Uppdelningen har gjorts med ett seed på 355. Mängden av både data är 333 observationer för träningsmängd respektive 167 observation för valideringsmängd.

Tabell 1: Frekvenstabell för spam på dataTrain

Spam	Freq
0	193
1	140

Träningsmängd innehåller 42.04% observation tillhör klass 1 och 57.96% observation tillhör klass 2.

Tabell 2: Frekvenstabell för spam på dataValid

Spam	Freq
0	97
1	70

Valideringsmängd innehåller 41.92% observation tillhör klass 1 och 58.08% observation tillhör klass 2. Klasserna är relativt balanserade i både mängderna med en något högre andel av klass 0 på omkring 58% jämfört med andel av klass 1 som utgör omkring 42% i båda.

2.2 b) Anpassa följande logistiska regressionsmodeller

2.2.1 i.alla förklarande variabler som mäter frekvensen av antalet ord (variabler döpta Word i data)

$$\text{logit}(P(Y=1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Word}_1 + \hat{\beta}_2 \cdot \text{Word}_2 + ... + \hat{\beta}_{48} \cdot \text{Word}_{48}$$

Där:
P är sannolikhet för spam ord.
$\hat{\beta}_0$ är interceptet.
$\hat{\beta}_1 - \hat{\beta}_{48}$ är koefficienter för respektive förklarande variabler (word1 - word48).

```
modell1 <- glm(formula = Spam ~.-Char1-Char2-Char3-Char4-Char5-Char6-Capitalrun1-
                 Capitalrun2-Capitalrun3,data = dataTrain,family = "binomial")
summary(modell1) %>%
  coef() %>%
  kable(digits = 3, caption = "Modellens skattade koefficienter.",
        col.names=c("Variabel", "Skattning", "Medelfel", "z-värde", "p-värde"))
```

Tabell 3: Modellens skattade koefficienter.

Variabel	Skattning	Medelfel	z-värde	p-värde
(Intercept)	1.007336e+14	8727436	11542173	0
Word1	-2.488191e+14	11601543	-21447068	0
Word2	-2.611610e+14	2620442	-99662978	0
Word3	1.293859e+14	9113105	14197781	0
Word4	1.361855e+14	1662367	81922641	0
Word5	2.477842e+14	6064200	40860154	0
Word6	7.967496e+14	17719066	44965666	0
Word7	8.115960e+13	13484150	6018889	0
Word8	3.725603e+14	9750781	38208249	0
Word9	5.755111e+14	15021156	38313368	0
Word10	2.512697e+14	7105490	35362761	0
Word11	-3.550899e+14	25742196	-13794079	0
Word12	-1.348723e+14	6061708	-22249877	0
Word13	-4.347919e+14	14981941	-29021068	0
Word14	2.512368e+14	20454185	12282902	0
Word15	5.480178e+14	24824499	22075685	0
Word16	3.301388e+14	4958541	66579836	0
Word17	6.016092e+14	8074325	74508920	0
Word18	2.363413e+14	8036650	29407939	0
Word19	-1.573590e+13	2622363	-6000656	0

Variabel	Skattning	Medelfel	z-värde	p-värde
Word20	9.622728e+14	7426644	129570338	0
Word21	-1.845593e+14	4287259	-43048325	0
Word22	1.248346e+14	3331105	37475421	0
Word23	8.458046e+14	9483686	89185218	0
Word24	1.697903e+14	6053344	28049014	0
Word25	-6.064802e+14	2798575	-216710408	0
Word26	-5.358414e+14	9066013	-59104419	0
Word27	-3.076256e+14	1512472	-203392571	0
Word28	2.050545e+14	11990797	17100993	0
Word29	-4.512479e+14	13644518	-33071736	0
Word30	-1.773967e+14	11932348	-14866871	0
Word31	1.541163e+15	36752100	41934018	0
Word32	9.066071e+14	43616546	20785853	0
Word33	-1.776267e+14	11427023	-15544441	0
Word35	-9.901469e+14	16831521	-58826943	0
Word36	6.789592e+14	22205943	30575563	0
Word37	8.443247e+13	10557748	7997204	0
Word38	-3.504347e+14	40563883	-8639083	0
Word39	-6.734460e+13	12061802	-5583295	0
Word40	-3.534903e+14	26400721	-13389420	0
Word41	-1.878989e+15	26022261	-72206988	0
Word42	-8.963136e+14	11840786	-75697137	0
Word43	-7.208534e+14	13902757	-51849674	0
Word44	-1.737261e+14	15259195	-11385014	0
Word45	-5.398892e+14	6224011	-86742977	0
Word46	1.368141e+14	10328967	13245675	0
Word47	-5.917559e+15	64813963	-91300689	0
Word48	-1.668033e+15	31417271	-53092866	0

2.2.2 ii.alla förklarande variabler som mäter frekvensen av antalet symboler (variabler döpta Char i data)

$$\text{logit}(P(Y=1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot Char_1 + \hat{\beta}_2 \cdot Char_2 + \hat{\beta}_3 \cdot Char_3 + \hat{\beta}_4 \cdot Char_4 + \hat{\beta}_5 \cdot Char_5 + \hat{\beta}_6 \cdot Char_6$$

Där:
P är sannolikhet för spam ord.
$\hat{\beta}_0$ är interceptet.
$\hat{\beta}_1 - \hat{\beta}_6$ är koefficienter för respektive förklarande variabler (Char1 - Char6).

```
modell2 <- glm(formula = Spam ~ Char1 + Char2 + Char3 + Char4 + Char5 + Char6,
               data = dataTrain, family = "binomial")
summary(modell2) %>%
  coef() %>%
  kable(digits = 3, caption = "Modellens skattade koefficienter.",
        col.names=c("Variabel", "Skattning", "Medelfel", "z-värde", "p-värde"))
```

Tabell 4: Modellens skattade koefficienter.

Variabel	Skattning	Medelfel	z-värde	p-värde
(Intercept)	-1.120	0.214	-5.231	0.000
Char1	-0.164	0.508	-0.322	0.747
Char2	-3.681	1.078	-3.414	0.001
Char3	1.386	2.871	0.483	0.629
Char4	3.503	0.607	5.770	0.000
Char5	7.902	1.603	4.929	0.000
Char6	2.000	1.173	1.705	0.088

2.2.3 iii.alla förklarande variabler i data

$$\text{logit}(P(\text{Spam} = 1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Word}_1 + \hat{\beta}_2 \cdot \text{Word}_2 + ... + \hat{\beta}_{48} \cdot \text{Word}_{48} + \hat{\beta}_{49} \cdot \text{Char}_1 + ... + \hat{\beta}_{54} \cdot \text{Char}_6 + \hat{\beta}_{55} \cdot \text{Capitalrun}_1 + \hat{\beta}_{56} \cdot \text{Capitalrun}_2 + \hat{\beta}_{57} \cdot \text{Capitalrun}_3$$

Där:
P är sannolikhet för spam ord.
$\hat{\beta}_0$ är interceptet.
$\hat{\beta}_1 - \hat{\beta}_{57}$ är koefficienter för alla förklarande variabler som innehålls i data.

Tabell 5: Modellens skattade koefficienter.

Variabel	Skattning	Medelfel	z-värde	p-värde
(Intercept)	-1.368496e+14	9649674.73	-14181787.3	0
Word1	-4.413109e+13	11681341.75	-3777913.1	0
Word2	-2.658775e+14	2645510.11	-100501429.6	0
Word3	2.511774e+14	9628765.02	26086144.8	0
Word4	1.334430e+14	1668562.79	79974815.8	0
Word5	-2.474257e+14	6129006.02	-40369630.3	0
Word6	5.576330e+14	17844406.76	31249733.8	0
Word7	2.519924e+14	13578098.98	18558737.1	0
Word8	2.552874e+14	9870992.52	25862380.3	0
Word9	3.767980e+14	15793936.20	23857128.2	0
Word10	1.071383e+14	7239503.66	14799119.3	0
Word11	-1.274574e+15	26110780.33	-48814072.2	0
Word12	-2.183655e+13	6140573.70	-3556109.2	0
Word13	-6.160612e+13	15315674.07	-4022423.2	0
Word14	2.657600e+14	24401070.73	10891323.4	0
Word15	5.593407e+14	28896701.37	19356557.7	0
Word16	2.561489e+14	4998099.14	51249266.0	0
Word17	6.596773e+14	8176665.63	80678029.6	0
Word18	-1.215196e+12	8304818.14	-146324.2	0
Word19	-4.097714e+12	2703164.39	-1515895.2	0
Word20	1.021557e+15	7667960.16	133224132.8	0
Word21	-8.961243e+13	4328856.76	-20701177.0	0
Word22	1.356225e+14	3869194.33	35051861.8	0
Word23	8.079454e+14	9906650.34	81555861.2	0
Word24	4.242097e+13	6067518.49	6991485.4	0
Word25	-5.160202e+14	2833404.50	-182120203.8	0
Word26	-1.063231e+14	9284616.69	-11451529.8	0
Word27	-2.651965e+14	1524161.64	-173995022.0	0
Word28	-2.320304e+14	12340382.65	-18802528.8	0
Word29	-4.110542e+14	13762449.56	-29867808.6	0

Variabel	Skattning	Medelfel	z-värde	p-värde
Word30	3.334060e+14	12100679.45	27552668.9	0
Word31	-5.264105e+14	36950737.87	-14246278.5	0
Word32	6.877663e+14	43922752.26	15658542.4	0
Word33	-7.406622e+13	11669472.36	-6347006.6	0
Word35	-2.918369e+14	16906015.16	-17262313.6	0
Word36	8.677129e+14	22347082.70	38828912.1	0
Word37	-2.891109e+13	10827315.98	-2670199.4	0
Word38	1.405651e+14	40801724.97	3445077.3	0
Word39	-3.071064e+14	12103780.04	-25372766.8	0
Word40	3.292079e+14	26553931.98	12397709.8	0
Word41	-8.893111e+14	26377070.95	-33715308.3	0
Word42	-6.702789e+14	11864177.78	-56496028.4	0
Word43	-5.594158e+14	14399664.47	-38849224.0	0
Word44	-3.065073e+14	15354155.49	-19962495.5	0
Word45	-3.644286e+14	6285751.99	-57976926.0	0
Word46	-2.430924e+14	10408970.68	-23354122.3	0
Word47	-4.594021e+15	64986428.94	-70691998.1	0
Word48	-3.091213e+15	31848192.50	-97060858.7	0
Char1	-4.124496e+14	16556740.21	-24911282.6	0
Char2	8.984323e+12	23115710.00	388667.4	0
Char3	-1.205184e+15	84996494.78	-14179221.7	0
Char4	4.876520e+14	6915659.02	70514178.4	0
Char5	1.102862e+15	15739390.91	70070200.9	0
Char6	-5.066937e+12	16052413.92	-315649.5	0
Capitalrun1	-1.402591e+12	159063.25	-8817817.9	0
Capitalrun2	8.646508e+11	59064.94	14638984.8	0
Capitalrun3	-1.926510e+11	13255.71	-14533438.1	0

2.3 c) Tolka parameterskattningen för den första variabeln i modell i. och ii.

För modellen 1: parameterskattningen för första variabeln är $-2,488 \cdot 10^{14}$ som tabellen ovan visar och det kan transformeras som $e^{-2,488191 \cdot 10^{14}}$, vilket är otroligt nära 0. Då man kan säga att när word1 öka med en enhet, så är oddset att mejlet är spam lika med 0 eller sannolikt att mejlet är spam går mot 0.

För modellen 2: parameterskattningen för första variabeln är -0.164 som tabellen ovan visar och det kan transformeras som $e^{-0.164}$, blir det 0.85. Då man kan säga att när Char1 öka med en enhet, så är oddset att mejlet är spam ca $\frac{1}{0.85} = 1.18$ gånger mindre.

2.4 d) Utgå från modell iii. och testa med ett test ifall modellen bör reduceras till modell i..

För att testa att modell iii (innehåll alla förklarande variabler) som kan reducera till modell i (innehåller alla förklarande variabler som mäter frekvensen av antalet ord) likelihoodkvottest beräknas:

$$H_0 : \beta_{49} = \beta_{50} = \beta_{51} = \beta_{52} = \beta_{53} = \beta_{54} = \beta_{55} = \beta_{56} = \beta_{57} = 0$$
$$H_a : \text{Minst en av } \beta_{49}, \beta_{50}, \beta_{51}, \beta_{52}, \beta_{53}, \beta_{54}, \beta_{55}, \beta_{56}, \beta_{57} \neq 0$$

Test formeln

$$LR_{test} = -2ln(\frac{L_R}{L_F}) = -2lnL_R - (-2lnL_F)$$

Med hjälper från R

```
LR <- (-2*logLik(modell1) - (-2*logLik(modell_full))) %>% c()
LR
```

```
[1] 1658.008
```

```
pValue <- pchisq(q = LR, df = 9, lower.tail = FALSE)
pValue
```

```
[1] 0
```

Med testet i träningsmängden fick vi att LR-värden är ganska stora och P-värden är mindre än signifikansnivån på 5%, vilket innebär att vi kan förkasta H_0. Detta innebär att den fullständiga modellen (som innehåller alla förklarande variabler) är bättre än modell i (som innehåller alla förklarande variabler som mäter antal ord) i träningsmängden.

2.5 e) Ta fram lämpliga utvärderingsmått på tränings- och valideringsmängden för de anpassade modellerna ur b). Presentera dessa i lämpliga tabeller för att underlätta jämförelsen. Kommentera och analysera era resultat. Vilken modell anser ni vara bäst för att modellera responsvariabeln och hur generaliserbar är modellerna?

2.5.1 Modell i - Träningsmängd

```
evaluation$confusionMatrix %>%
  kable(caption = "Modellens förväxlingsmatris")
```

Tabell 6: Modellens förväxlingsmatris

	0	1
0	153	40
1	16	124

Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 0$ är 153.
Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 1$ är 40.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 0$ är 16.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 1$ är 124.

```
evaluation$overall %>%
  kable(caption = "Modellens övergripande mått", col.names = c("Träffsäkerhet",
                                                               "Felkvot"),
        digits = 3)
```

Tabell 7: Modellens övergripande mått

Träffsäkerhet	Felkvot
0.832	0.168

Med en träffsäkerhet på 0.832 innebär det att modellen korrekt klassificerade 83.2% av alla observationer. En felkot på 0.168 innebär att modellen gjorde fel på 16.8% av alla observationer.

```
evaluation$classWise %>%
  kable(caption = "Modellens klasspecifika mått", digits = 3)
```

Tabell 8: Modellens klasspecifika mått

	0	1
sensitivitet	0.793	0.886
specificitet	0.886	0.793

För klass 0, en sensitivitet (0.793) och en högre specificitet (0.886), vilket innebär att modellen är bättre på att hitta faktiska som inte klass 0-fall än att korrekt identifiera fall som tillhör klass 0. För klass 1, en hög sensitivitet (0.886) och lägre specificitet (0.793), vilket innebär att modellen är sämre på att hitta faktiska som inte klass 1-fall än på att korrekt identifiera fall som tillhör klass 1.

2.5.2 Modell i - Valideringsmängd

```
evaluation$confusionMatrix %>%
  kable(caption = "Modellens förväxlingsmatris")
```

Tabell 9: Modellens förväxlingsmatris

	0	1
0	77	20
1	10	60

Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 0$ är 77.
Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 1$ är 20.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 0$ är 10.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 1$ är 60.

```
evaluation$overall %>%
  kable(caption = "Modellens övergripande mått", col.names = c("Träffsäkerhet",
                                                               "Felkvot"),
        digits = 3)
```

Tabell 10: Modellens övergripande mått

Träffsäkerhet	Felkvot
0.82	0.18

Med en träffsäkerhet på 0.82 innebär det att modellen korrekt klassificerade 82% av alla observationer. En felkot på 0.18 innebär att modellen gjorde fel på 18% av alla observationer.

```
evaluation$classWise %>%
  kable(caption = "Modellens klasspecifika mått", digits = 3)
```

Tabell 11: Modellens klasspecifika mått

	0	1
sensitivitet	0.794	0.857
specificitet	0.857	0.794

För klass 0, en sensitivitet (0.794) och en högre specificitet (0.857), vilket innebär att modellen är bättre på att hitta faktiska som inte klass 0-fall än på att korrekt identifiera fall som tillhör klass 0. För klass 1, en hög sensitivitet (0.857) och lägre specificitet (0.794), vilket innebär att modellen är sämre på att faktiska som inte klass 1-fall än på att korrekt identifiera fall som tillhör klass 1.

2.5.3 Modell ii - Träningsmängd

```
evaluation$confusionMatrix %>%
  kable(caption = "Modellens förväxlingsmatris")
```

Tabell 12: Modellens förväxlingsmatris

	0	1
0	175	18
1	41	99

Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 0$ är 175.
Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 1$ är 18.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 0$ är 41.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 1$ är 99.

```
evaluation$overall %>%
  kable(caption = "Modellens övergripande mått", col.names = c("Träffsäkerhet",
                                                               "Felkvot"),
        digits = 3)
```

Tabell 13: Modellens övergripande mått

Träffsäkerhet	Felkvot
0.823	0.177

Med en träffsäkerhet på 0.823 innebär det att modellen korrekt klassificerade 82.3% av alla observationer. En felkot på 0.177 innebär att modellen gjorde fel på 17.7% av alla observationer.

```
evaluation$classWise %>%
  kable(caption = "Modellens klasspecifika mått", digits = 3)
```

Tabell 14: Modellens klasspecifika mått

	0	1
sensitivitet	0.907	0.707
specificitet	0.707	0.907

För klass 0, en hög sensitivitet (0.907) och en lägre specificitet (0.707), vilket innebär att modellen är bättre på att hitta faktiska klass 0-fall än på att korrekt identifiera fall som inte tillhör klass 0.
För klass 1, en sensitivitet (0.707) och högre specificitet (0.907), vilket innebär att modellen är sämre på att hitta faktiska klass 1-fall än på att korrekt identifiera fall som inte tillhör klass 1.

2.5.4 Modell ii - Valideringsmängd

```
evaluation$confusionMatrix %>%
  kable(caption = "Modellens förväxlingsmatris")
```

Tabell 15: Modellens förväxlingsmatris

	0	1
0	88	9
1	31	39

Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 0$ är 88.
Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 1$ är 9.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 0$ är 31.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 1$ är 39.

```r
evaluation$overall %>%
  kable(caption = "Modellens övergripande mått", col.names = c("Träffsäkerhet",
                                                               "Felkvot"),
        digits = 3)
```

Tabell 16: Modellens övergripande mått

Träffsäkerhet	Felkvot
0.76	0.24

Med en träffsäkerhet på 0.76 innebär det att modellen korrekt klassificerade 76% av alla observationer. En felkot på 0.24 innebär att modellen gjorde fel på 24% av alla observationer.

```r
evaluation$classWise %>%
  kable(caption = "Modellens klasspecifika mått", digits = 3)
```

Tabell 17: Modellens klasspecifika mått

	0	1
sensitivitet	0.907	0.557
specificitet	0.557	0.907

För klass 0, en hög sensitivitet (0.907) och en lägre specificitet (0.557), vilket innebär att modellen är bättre på att hitta faktiska klass 0-fall än på att korrekt identifiera fall som inte tillhör klass 0.
För klass 1, en sensitivitet (0.557) och högre specificitet (0.907), vilket innebär att modellen är sämre än på att faktiska klass 1-fall än på att korrekt identifiera fall som inte tillhör klass 1.

2.5.5 Modell iii - Träningsmängd

```r
evaluation$confusionMatrix %>%
  kable(caption = "Modellens förväxlingsmatris")
```

Tabell 18: Modellens förväxlingsmatris

	0	1
0	183	10

	0	1
1	23	117

Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 0$ är 183.
Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 1$ är 10.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 0$ är 23.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 1$ är 117.

```
evaluation$overall %>%
  kable(caption = "Modellens övergripande mått", col.names = c("Träffsäkerhet",
                                                               "Felkvot"),
        digits = 3)
```

Tabell 19: Modellens övergripande mått

Träffsäkerhet	Felkvot
0.901	0.099

Med en träffsäkerhet på 0.901 innebär det att modellen korrekt klassificerade 90.1% av alla observationer.
En felkot på 0.099 innebär att modellen gjorde fel på 9.9% av alla observationer.

```
evaluation$classWise %>%
  kable(caption = "Modellens klasspecifika mått", digits = 3)
```

Tabell 20: Modellens klasspecifika mått

	0	1
sensitivitet	0.948	0.836
specificitet	0.836	0.948

För klass 0, en hög sensitivitet (0.948) och en lägre specificitet (0.836), vilket innebär att modellen är bättre på att hitta faktiska klass 0-fall än på att korrekt identifiera fall som inte tillhör klass 0.
För klass 1, en sensitivitet (0.836) och högre specificitet (0.948), vilket innebär att modellen är sämre på att faktiska klass 1-fall än på att korrekt identifiera fall som inte tillhör klass 1.

2.5.6 Modell iii - Valideringsmängd

```
evaluation$confusionMatrix %>%
  kable(caption = "Modellens förväxlingsmatris")
```

Tabell 21: Modellens förväxlingsmatris

	0	1
0	82	15
1	14	56

Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 0$ är 82.
Frekvensen av antalet Y = 0 som blivit predikterade till klass $\hat{Y} = 1$ är 15.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 0$ är 14.
Frekvensen av antalet Y = 1 som blivit predikterade till klass $\hat{Y} = 1$ är 56.

```
evaluation$overall %>%
  kable(caption = "Modellens övergripande mått", col.names = c("Träffsäkerhet",
                                                               "Felkvot"),
        digits = 3)
```

Tabell 22: Modellens övergripande mått

Träffsäkerhet	Felkvot
0.826	0.174

Med en träffsäkerhet på 0.826 innebär det att modellen korrekt klassificerade 82.6% av alla observationer. En felkot på 0.174 innebär att modellen gjorde fel på 17.4% av alla observationer.

```
evaluation$classWise %>%
  kable(caption = "Modellens klasspecifika mått", digits = 3)
```

Tabell 23: Modellens klasspecifika mått

	0	1
sensitivitet	0.845	0.800
specificitet	0.800	0.845

För klass 0, en hög sensitivitet (0.845) och en lägre specificitet (0.800), vilket innebär att modellen är bättre på att hitta faktiska klass 0-fall än på att korrekt identifiera fall som inte tillhör klass 0.
För klass 1, en sensitivitet (0.800) och högre specificitet (0.845), vilket innebär att modellen är sämre på att faktiska klass 1-fall än på att korrekt identifiera fall som inte tillhör klass 1.

2.5.7 Slutsat

Enligt uppgiften bedöms modell iii vara bättre än modell i, men detta gäller endast för träningsmängden, vilket syns tydligt genom träffsäkerheten på 90,7% jämfört med 82,3%. Det betyder dock inte att modell iii är bäst för att modellera responsvariabeln eller har den högsta generaliserbarheten. De klasspecifika måtten ligger jämnt mellan 80–90% för både träningsmängden och valideringsmängden för både modell i och iii, vilket antyder att båda modellerna är lika bra på att prediktera klass 1 som klass 0. För modell iii är skillnaden i träffsäkerhet mellan träningsmängden och valideringsmängden relativt stor (90,7% jämfört med 82,6%). Detta är en större än skillnaden i träffsäkerhet mellan träningsmängden och valideringsmängden för modell i, vilket indikerar att modell i kan ge mer stabila värden även med olika data. För modell ii är träffsäkerheten för träningsmängden relativt hög, nämligen 82,3%, men den sjunker till 76% för valideringsmängden, vilket indikerar att modellen överanpassar. Dessutom är modell ii inte lika bra på att prediktera klass 1 som klass 0; för träningsmängden är sensitiviteten 0,707 jämfört med 0,907, och för valideringsmängden är den 0,557 jämfört med 0,907.

Sammanfattningsvis kan vi dra slutsatsen att modell I är bäst och mest generaliserbar.

3 Lärdomar, problem, övriga kommentarer

Efter vi utförde arbetet har vi fått grundläggande koll på hur logiska regression fungerar/tillämpas i samband med träning-valideringsmängder för modells anpassning och utvärdering.