Överlevnadsanalys av Titanicolyckan

732G39

Michael Debebe
Sandra Vesterling

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2022-06-01

Handledare: Karl Wahlin, statistikkonsult
Examinator: Isak Hietala, universitetsadjunkt

Sammandrag

Denna uppsats analyserar ett datamaterial med information om passagerarna ombord fartyget Titanic. Inledningsvis studeras datamaterialet, de moment som används för binär logistisk regression och hur resultaten mäts. Begreppen logit, oddskvot och sannolikhet förtydligas, även en fördjupning av matematiken bakom den logistiska funktionen presenteras. Den binära logistiska regressionen används idag i stor utsträckning och lämpar sig mycket väl för responsvariablar med två utfall.

Den slutliga logistiska regressionsmodellen består av 4 parametrar varav 1 kommer från en interaktion mellan ålder och kön. Utifrån denna modell analyserades variablernas påverkan på responsvariabeln.

Resultatet blev att det finns en skillnad mellan både kön och/eller olika åldrar. Kvinnor och barn hade en större chans än vuxna män att överleva olyckan.

Abstract

This thesis analyzes a data material with information about the passengers onboard the ship Titanic. It begins by studying the data material, the steps used for binary logistic regression and how the results are measured. The concepts of logit, odds ratio and probability are clarified, there will also be an in-depth study of the mathematics behind the logistics function. The binary logistic regression is widely used today and is very well suited for response variables with two outcomes.

The final logistic regression model consists of four parameters, one of which come from an interaction between age and gender. Based on this model, the impact of the variables on the response variable was analyzed.

The result was that there is a difference between all genders and/or different ages. Women and children had a greater chance than adult men of surviving the accident.

Innehåll

Tabeller

Figurer

1. Introduktion

I rapportens första kapitel ges en bakgrund till ämnet, formulering av syfte och problemställningar samt genomgång av etiska och samhälleliga aspekter.

1.1 Bakgrund

10:e april 1912 satte Titanic i väg på sin jungfruresa, fyra dagar in på resan krockade fartyget med ett isberg och sjönk utanför Kanada den 15 april 1912. Det var ca 70 procent av passagerarna och besättningen som dog i olyckan. Ombord fartyget fanns det 1178 platser i livbåtarna trots att fartyget var planerat för sammanlagt 3300 personer. Faktumet att det fanns färre platser på livbåtarna än passagerare orsakade förvirring, vilket ledde till att enbart hälften av livbåtarnas 1178 platser användes, resultatet blev att drygt 70 % av människorna ombord omkom (Ulfvarson, 2022).

Det råder en allmän uppfattning om att kvinnor och barn är de som går först i händelse av kris, det syns såväl i dåtid som i nutid. Därav har gruppen kommit fram till att studera överlevnadschansen baserat på kön och ålder som huvudvariabler med interaktionsterm mellan dem. Tidigare studier ser en möjlig förklaring till att kvinnor och barn hade störst överlevnadsgrad på just Titanic, utifrån att kaptenen beordrade ”kvinnor och barn först”, men att kvinnor och barn annars har en avsevärt lägre överlevnadsgrad vid fartygsolyckor. (Elinder och Erixson, 2012).

1.2 Syfte

Syftet med denna rapport är att undersöka vilka de bidragande orsakerna till att man överlevde olyckan var. Då gruppens huvudfokus ligger på variablerna kön och ålder resulterade frågeställningarna i de som presenteras nedan.

- Hur skiljer sig sannolikheten att överleva könen emellan?
- Hur skiljer sig sannolikheten att överleva baserat på ålder?
- Hur överlevnadschansen påverkas av att huvudvariablerna interagerar med varandra?

1.3 Etiska och samhälleliga aspekter

De etiska aspekter som kan identifieras är att det år 1912, förmodligen fanns stora skillnader samhällsklasser emellan, vilket blir tydligt i fartygets olika reseklasser. Detta arbete kan komma att ta upp är hur stor påverkan en persons ekonomiska och sociala tillhörighet hade på chansen att överleva olyckan. Detta kan ge en större inblick i konsekvenserna av ett klassamhälle. Faktumet att studien syftar till att analysera huruvida social status har en påverkan på överlevnadsgrad är ett känsligt ämne, detta i samband med att värde av liv spekuleras.

2. Data

I rapportens andra kapitel ges en beskrivning av variablerna, dess transformationer samt beskrivande statistik över datamaterialet.

Datamaterialet är hämtat från Stanfords Universitys hemsida och innehåller 887 observationer där varje rad motsvarar en person. Datamaterialet har åtta kolumner där den första är den binära responsvariabeln och sedan sju förklarande variabler varav två som kommer användas i denna analys (Stanford University, 2016). Datamaterialet har behandlats i R Studio.

2.1 Beskrivning av variabler

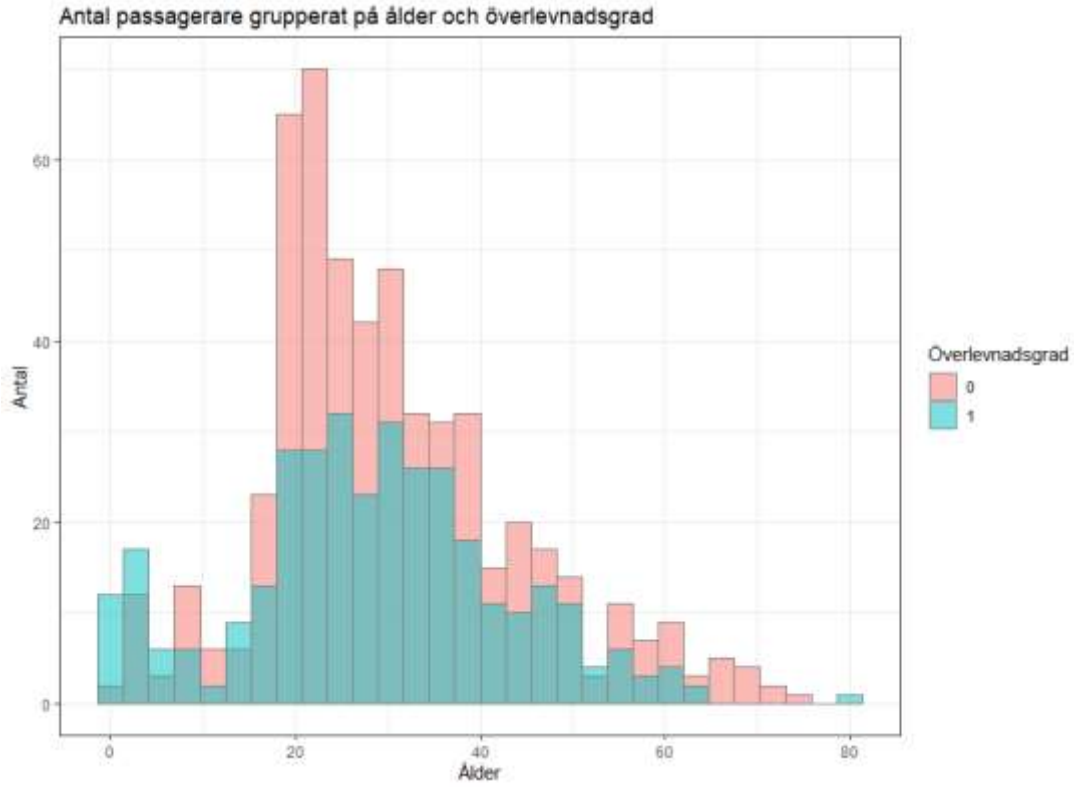
I detta avsnitt ges en beskrivning av de olika variablerna i datamaterialet.

Tabell 1:Beskrivningstabell över variabler

Variabelnamn	**Beskrivning**
Survived Indicator	0 = Personen överlevde ej 1 = Personen överlevde
Sex	Personens kön
Age	Personens ålder

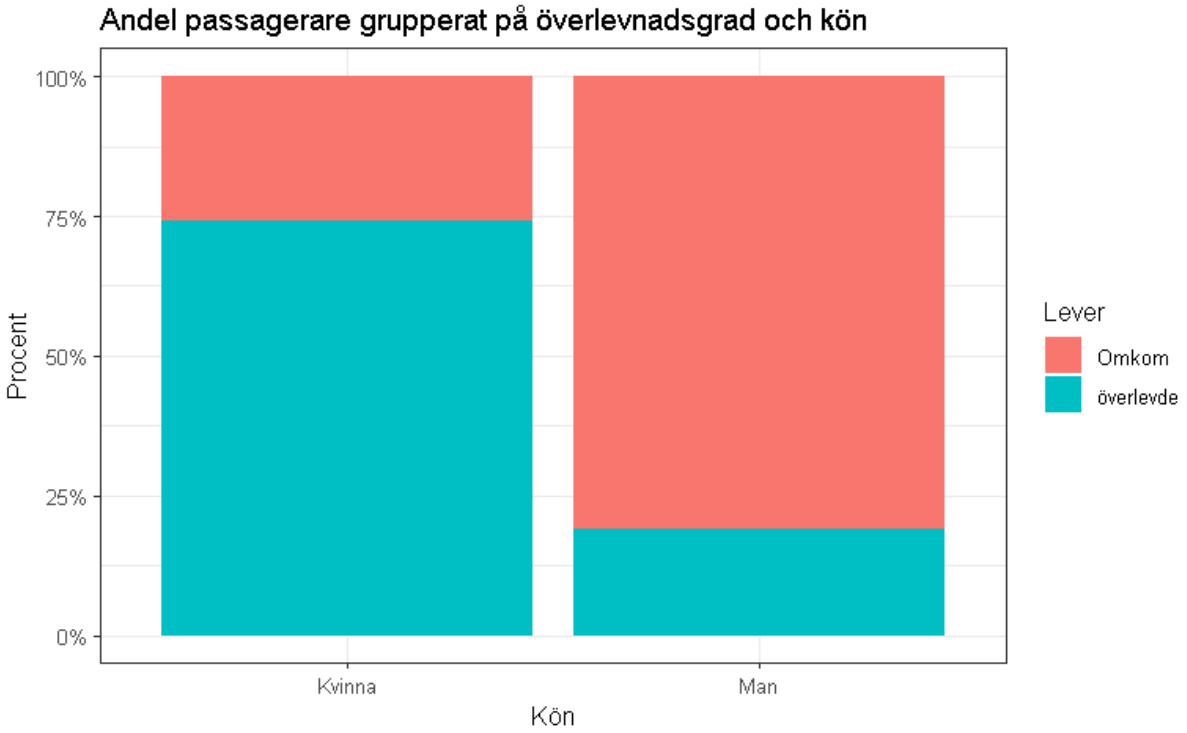
2.3 Beskrivande statistik

I detta avsnitt visas beskrivande statistik för överlevnadsgraden fördelat på de variabler som avses analyseras i denna rapport.



Figur 1: Antal passagerare grupperat på ålder och överlevnadsgrad

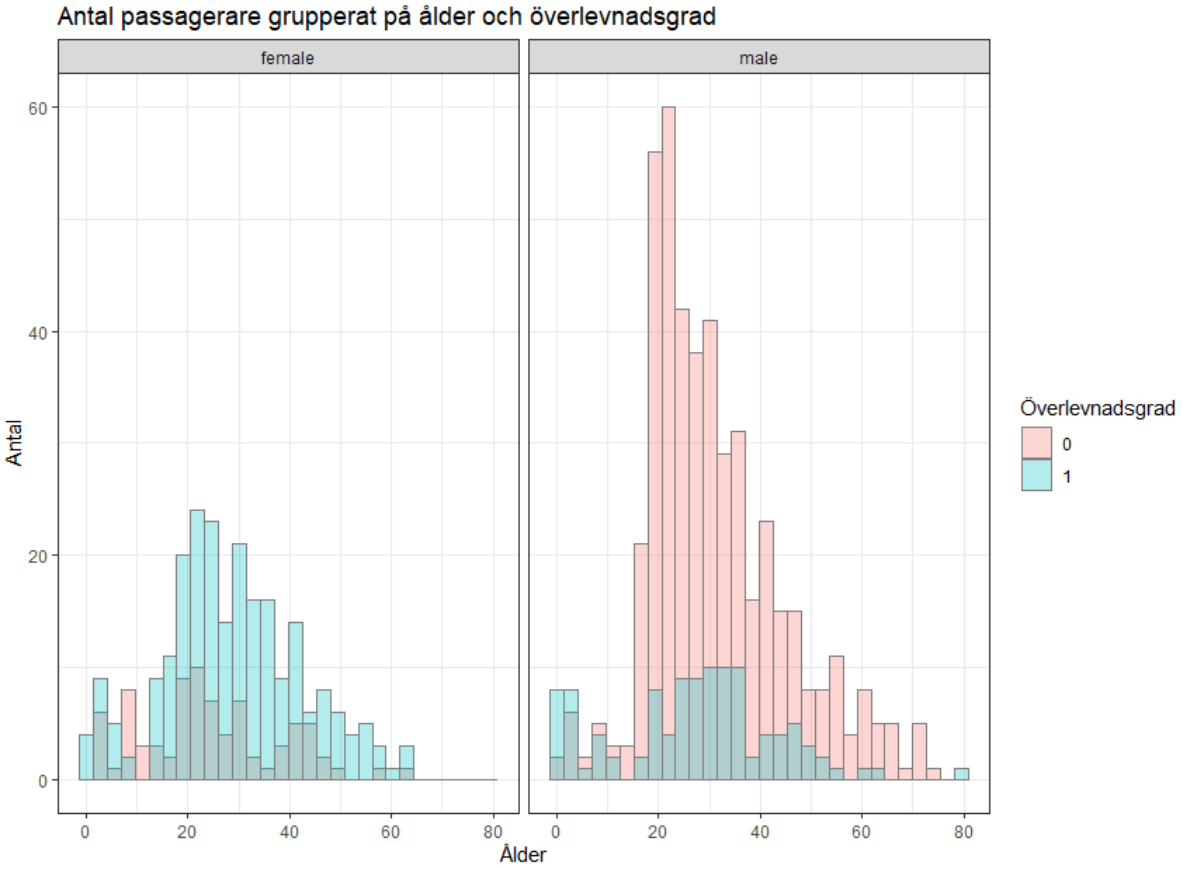
I figur 1 visas antal överlevande och omkomna för varje ålder. Majoriteten av de yngre åldrarna (0—15 år) har ett högre antal överlevande än omkomna medan för de högre åldrarna (20–50 år) är det tvärtom. I samband med att åldern ökar syns det hur förhållandet mellan överlevnadsgraden jämnar ut sig inom varje ålder jämfört med åldrarna i mitten.



Figur 2: Andel passagerare grupperat på överlevnadsgrad och kön

I figur 2 visas andelen överlevande och omkomna för vardera kön. Det var en större andel av männen som omkom än som överlevde och för kvinnorna var resultatet det omvända.

Detta kan kopplas ihop med det som nämns i (**Bakgrund**), vad gäller att kvinnor hade större överlevnadsgrad ombord Titanic, vilket syns ovan, där totalt 74,2 % av alla kvinnor överlevde medan enbart 19,1 % av männen överlevde.



Figur 3: Antal passagerare grupperat på kön, ålder och överlevnadsgrad

I figur 3 syns antalet överlevande och omkomna fördelat på både kön och ålder. För de flesta åldrar var det fler kvinnor som överlevde än som omkom och för männen var det tvärtom. För män är enbart de yngsta åldrarna samt de äldsta av männen med fler överlevande än omkomna.

2.4 Datauppdelning

I samband med att en logistisk regressionsmodell utformas, kan datauppdelning användas, där det insamlade materialet delas upp till en tränings-, validerings- och/eller testmängd för att motverka överanpassning vid övervakad inlärning (Hietala, 2020). Datauppdelning sker alltså i syfte att motverka överanpassning samt kunna validera modellen på testdata då denne tränats på träningsdata. Metoden är användbar då det finns sex till tio gånger så många observationer som det finns förklarande variabler (H.Kutner, 2005).

3. Metod

I rapportens tredje kapitel ges en förklaring av de modeller, formler och mått som används senare i rapporten för att undersöka datamaterialet och besvara de givna frågeställningarna.

3. 1 Logistisk regression

För detta datamaterial kommer binär logistisk regression användas. Binär logistisk regression används när responsvariabeln kan anta ett av två möjliga värden (Klienbaum, 2021).

Den logistiska regressionsmodellen ser ut som följer:

$$E[Y] = \hat{\pi} = \frac{1}{1 + \exp\left[-(\beta_0 + \sum_{i=1}^{k} \beta_i X_i)\right]} \quad (1)$$

Där X_i är de förklarande variablerna, $i = 1, 2 \ldots, k$, och $\beta -$ parametrarna i modellen skattas med maximum likelihood-metoden som hittar det mest troliga värdet på parametern baserat på observationerna som finns i datamaterialet. För den logistiska regressionsmodellen används 0,5 som ett gränsvärde, där Y predikteras till 1 om $\hat{\pi} > 0{,}5$ och 0 annars.

Likelihood funktionen ser ut som följer:

$$L(\beta_i) = \prod_{i=1}^{n} \pi = \prod_{i=1}^{n} \frac{1}{1 + \exp\left[-(\beta_0 + \sum_{i=1}^{k} \beta_i X_i)\right]} \quad (2)$$

För att transformera $E[Y]$ till Y behövs en länkfunktion och med hjälp av logit-länken ges:

$$logit(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i X_i \quad (3)$$

Därefter kan oddskvoter definieras som är ett mått för de skattade $\beta -$ parametrarnas effekt på responsvariabeln, givet att resterande hålls fixa. Oddskvoten fås från kvoten av sannolikheten av en händelse dividerat med komplementet av sannolikheten till den händelsen. Där en oddskvot större än 1 betyder att sannolikheten för händelsen är större och odds lägre än 1 betyder att sannolikheten är lägre (Klienbaum, 2021):

$$OR_A = \frac{P(A)}{1 - P(A)} \quad (4)$$

Oddskvoten tolkas som en förändringsfaktor.

3.1.1 Wald-test

För att undersöka signifikanser för enskilda parametrar i modellen kan Wald-test utföras. Detta hjälper till att avgöra om variabeln bör inkluderas i modellen eller inte. För detta ställs följande hypoteser upp där variabeln bör inkluderas om H_0 kan förkastas (Klienbaum, 2021):

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

Testet beräknas genom:

$$W = \frac{\mathrm{b}_i}{S_{b_i}} \quad (5)$$

Där $\mathrm{b_i}$ är parametern som skattas och S_{b_i} är parameterns standardavvikelse. H_0 kan förkastas om $|W| > Z\left(1 - \frac{\alpha}{2}\right)$.

3.4 Modellutvärdering

För vidare analys behöver modellen undersökas med hjälp av nedanstående metoder.

3.4.1 Förväxlingsmatris

		True (Observed) Outcome	
		Y = 1	Y = 0
Predicted	Y = 1	$n_{TP} = 70$	20
Outcome	Y = 0	30	$n_{TN} = 80$
		$n_1 = 100$	$n_0 = 100$

Figur 4: Förväxlingsmatris

Inom den logistiska regressionen presenteras modellens precision gällande prediktioner med hjälp av förväxlingsmatris, observerade och predikterade utfall kombineras i en tabell, likt den som visas ovan. Matrisen fokuserar på två kvantiteter, antalet sanna utfall och antalet felaktiga utfall, prediktionerna jämförs alltså mot det sanna utfallet. Det sanna utfallet består av (Sann negativ och Sann positiv) medan de felaktiga utfallet består av (Falsk negativ och falsk positiv) (David G. Kleinbaum, 2010).

3.4.2 Felkvot

Felkvoten kännetecknar hur väl modellen predikterar nya observationer. Felkvoten för en modell är förhållandet mellan antalet felaktiga enheter i modellen och totala antalet enheter (David G. Kleinbaum, 2010).

$$Felkvot = \frac{Felaktiga\ enheter}{\text{Totala antalet enheter}} \quad (6)$$

3.4.3 Sensitivitet

Sensitivitet mäter andelen faktiska positiva resultat som korrekt identifieras med andra positiva resultat och kompletterar den falska negativa frekvensen, även känt som ”type 2 Error” (David G. Kleinbaum, 2010).

$$Sensitivitet = \frac{Sann\ positiv}{Faktiskt\ sanna\ (Sann\ positiv + Sann\ negativ)} \quad (7)$$

3.4.4 Specifitivitet

Specifitivitet mäter till skillnad från **3.4.3 Sensitivitet** andelen sanna negativa utfall och kan räknas som 1-andelen falsk negativa utfall, även känt som ”type 1 Error” (David G. Kleinbaum, 2010).

$$Specifitivitet = \frac{Sann\ negativ}{Sann\ negativ + Falsk\ positiv} \tag{8}$$

3.4.5 VIF-värden

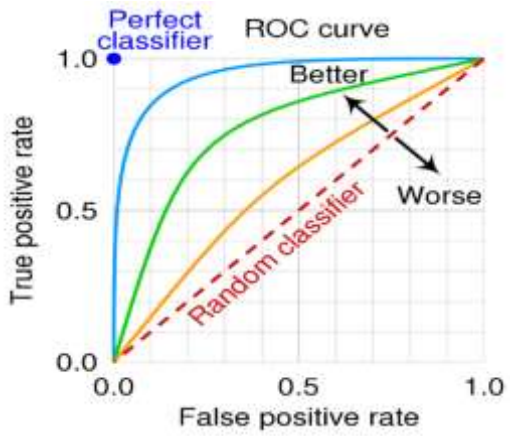
Om förklarande variabler har ett starkt samband mellan sig, riskerar det att uppstå problem med multikollinearitet. För att identifiera multikollinearitet används VIF (Variance inflation factor).

$$VIF_j = \frac{1}{1 - R_j^2} \tag{9}$$

Där $j > 0$ och R_j^2 beräknas från en regressionsmodell med X_j som den angivna förklarande variabeln. Tröskeln för VIF_j är subjektiva men ska inte vara för stora (> 10). Om stora VIF_j erhålls kan problemet med multikollinearitet lösas genom att ta bort påverkande variabler. (Klienbaum, 2021).

3.4.6 ROC-kurvor

ROC-kurvan är en grafkurva som visualiserar den diagnostiska förmågan hos ett binärt klassificeringssystem efter att dess tröskelvärde för sannolikhetsskattningar varieras. ROC-kurvan skapas genom att plotta den sanna positiva frekvensen (**3.4.3 Sensitivitet**), gentemot den falska positiva frekvensen vid olika threshold värden. ROC-kurvan är sammanfattningsvis ett bra verktyg för att välja de optimala modellerna och förkasta de suboptimala oberoende från kostnad eller klassfördelning (David G. Kleinbaum, 2010).



Figur 5: ROC-KURVA (Thoma, 2018)

Den diagnostiska förmågan mäts utifrån att ju närmare kurvan går det övre vänstra hörnet desto bättre är den diagnostiska förmågan. För en perfekt modell går kurvan från (0,0) till (0,1) och sedan diagonalt till punkten (1,1) (SBU, 2020).

4. Resultat

I det fjärde kapitlet presenteras den logistiska regressionsmodellen som anpassats, analys av variablerna, utvärdering av modellen samt prediktioner.

4.1 Logistiska regressionsmodellen

I tabell 2 nedan visas resultaten från den logistiska regressionsmodellen med variablerna kön och ålder samt en interaktionsterm mellan dessa. Den slutgiltiga modellen innehåller 4 parametrar där referensgruppen är kvinnor. Indikatorvariabeln Sexmale finns för att kännetecknа män.

Tabell 2: Modellen

Parameter	**Skattning**	**Std.Error**	**Z-value**	**Pr(>|z|)**
(Intercept)	**0.4247**	**0.3318**	**1.28**	**0.20061**
Sexmale	**-1.5885**	**0.4432**	**-3.58**	**0.00034 *** **
Age	**0.0212**	**0.0114**	**1.86**	**0.06323**
Sexmale: Age	**-0.0311**	**0.0146**	**-2.14**	**0.03262 ***

4.1.1 Parametrarnas signifikans

I kolumnen längst till höger i tabell 2 syns parametrarnas p-värde för det Wald-test som utförts för att undersöka enskilda parametrars signifikanser på signifikansnivå $\alpha = 0.05$ (**3.1.1 Wald-test**). För parametrarna Sexmale och Sexmale:Age kan nollhypotesen förkastas och dessa variabler har ett signifikant samband med responsvariabeln. Däremot har inte parametern Age själv ett signifikant samband med responsvariabeln på 5 % signifikansnivå men bör ändå vara med i modellen då den har en påverkan i interaktionen med kön.

4.1.2 VIF-värden

I tabell 3 nedan visas VIF-värdena för parametrarna som ingår i modellen.

Tabell 3: VIF-värden

	Frihetsgrader	VIF
Parameter		
Sex	1	4.97
Age	1	2.64
Sex: Age	1	7.25

Låga VIF-värden för samtliga variabler returneras och det verkar därmed inte finnas några problem med multikollinearitet i modellen (**3.4.5 VIF-värden**).

4.1.3 Odds-kvoter för modellen

För att analysera huvudvariablernas (kön och ålder) effekt på sannolikheten att överleva har oddskvoter (**3. 1 Logistisk regression**) beräknats. Dessa tolkas som en förändringsfaktor och hjälper att se vilka grupper som har större respektive mindre chans att överleva.

Tabell 4: Koefficienter med oddskvot

Koefficient	Odds-kvot
(Intercept)	1.525
Sexmale	0.205
Age	1.022
Sexmale: Age	0.969

I tabell 4 syns samtliga variabler och dess oddskvot. Vid en första anblick framgår det att kvinnor har en högre sannolikhet att överleva än att omkomma, detta då referensgruppen (interceptet), har 52,5 % större chans att överleva än att omkomma.

Indikatorvariabeln Sexmale har en oddskvot på 0,205, alltså har män 79,5 % mindre chans att överleva givet att resterande variabler hålls fixa. För variabeln ålder med en oddskvot på 1,022, ökar sannolikheten att överleva med 2,2 % för varje år passageraren blir äldre, givet att resterande variabler hålls fixa.

Sexmale:Age representerar oddskvoten för en manlig passagerare, för varje år äldre passageraren blir sjunker sannolikheten att överleva med 3,1 %.

4.2 Modellutvärdering

Då enbart en modell har anpassats görs en utvärdering av denna för att undersöka om resultaten är tillförlitliga.

4.2.1 Förväxlingsmatris

I detta delkapitel kommer validering av modellen som skapats på träningsdata att genomföras i form av jämförelser av precision mellan träningsdata och testdata.

Tabell 5: Förväxlingsmatris utvärderad på testdata

Observerat	**Predikterat**	
	0	1
0	119	29
1	23	77

Utifrån från denna matris kan fler utvärderingsmått beräknas. Modellen har en missklassificeringsgrad på 21%, alltså är 21% av modellens prediktioner felaktiga (**3.4.2 Felkvot**). Sensitiviteten summerar till 72,6 % vilket innebär att chansen för att modellen predikterar en passagerare till att överleva när den i verkliga fallet gjorde det är 72,6 % (**3.4.3 Sensitivitet**). Modellen har 83,8 % Specifitivitet vilket innebär att chansen för att modellen predikterar en passagerare till att överleva trots att den inte gjorde det är 16,2 % (**3.4.4 Specifitivitet**).

Tabell 6: Förväxlingsmatris utvärderad på träningsdata

Observerat	**Predikterat**	
	0	1
0	345	80
1	58	156

Även för denna matris beräknas det flera utvärderingsmått. Modellen har en missklassificeringsgrad på 21,6%, alltså är 0,6 procentenheter fler enheter av modellens prediktioner felaktiga gentemot testdata. Sensitiviteten summerar till 66,1 % vilket är 6,5 procentenheter högre än för testdata. Modellen har 85,6 % Specifitivitet vilket innebär att chansen för att modellen predikterar en passagerare till att överleva trots att den inte gjorde det är 14,4 %, alltså är Specifitiviteten 2,8 procentenheter lägre för träningsdata gentemot test.

4.2.2 Sannolikhetsberäkningar

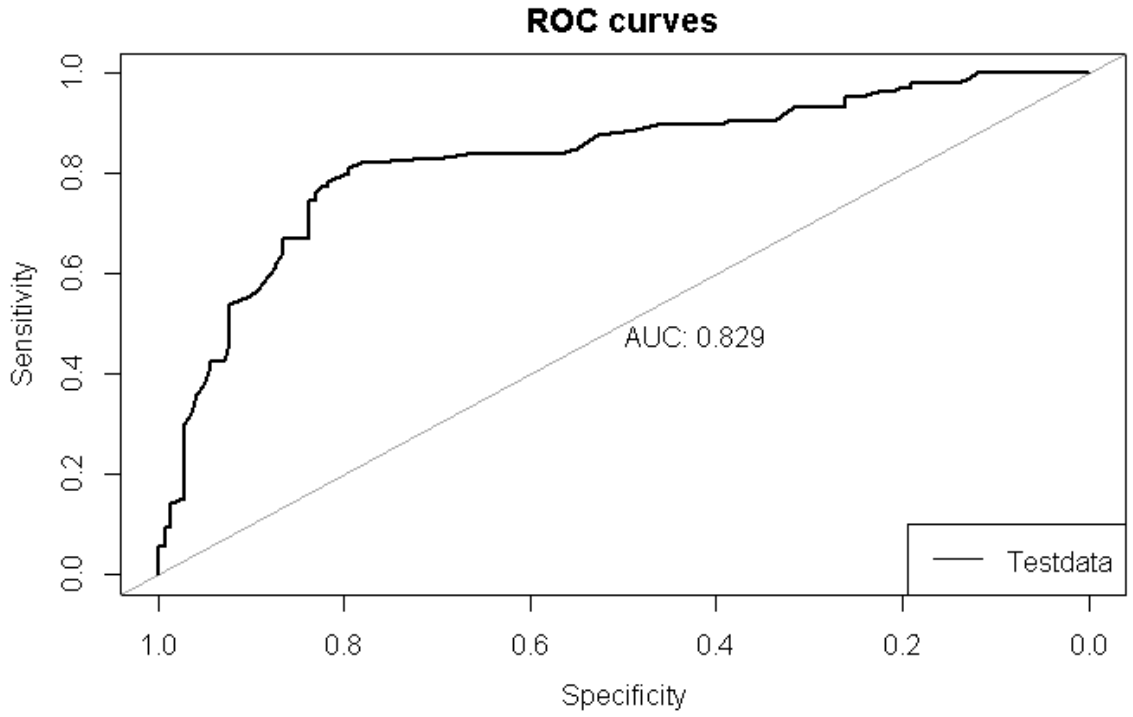
För att tydligare titta på hur modellen presterar visas sannolikhetsberäkningar och prediktioner för ett slumpat urval av 6 passagerare, i 2 av 6 fall har modellen predikterat fel.

Tabell 7: Sannolikhetsberäkningar för ett slumpat urval av 6 passagerare

Passagerare	Kön	Ålder	Sannolikhet att överleva	Predikterat	Utfall
1	Kvinna	16	0,682	Överlevde	Omkom
2	Man	25	0,196	Omkom	Överlevde
3	Kvinna	48	0,809	Överlevde	Överlevde
4	Man	4	.231	Omkom	Omkom
5	Man	11	.219	Omkom	Omkom
6	Kvinna	52	0,822	Omkom	Omkom

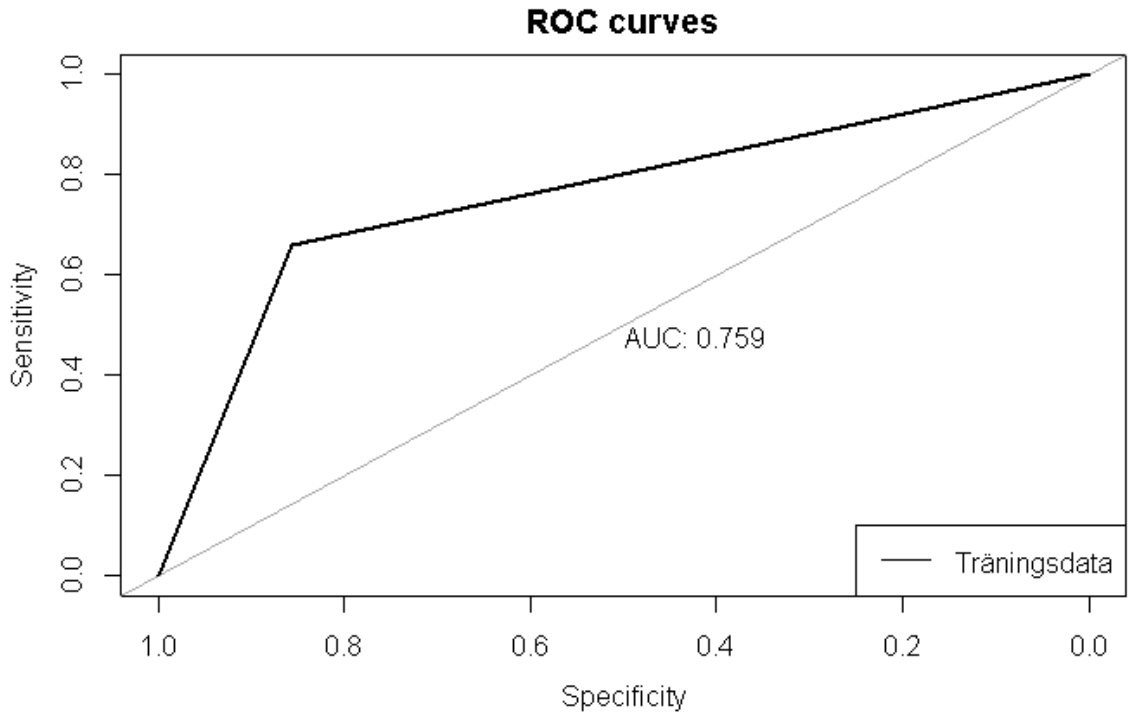
Modellen har predikterat fel för en kvinna som är 16 år och en man som är 25 år. Sannolikheten att överleva är betydligt högre för kvinnorna i detta urval än för männen. Kvinnorna har mellan 68,2% och 82,2% chans att överleva medan männen ligger mellan 19,6% och 23,1%.

4.2.3 ROC-kurvor



Figur 6:Roc-kurva för testdata

I figur 6 som visas ovan, representeras de tröskelvärdet som enligt Roc-kurvan returnerar så hög sensitivitet som möjligt samtidigt som att specifitiviteten är så hög som möjligt. Tröskelvärdet är som tidigare nämnt (1-AUROC), vilket för testdatat är ett tröskelvärde på 0,171, som för detta datamaterial är väldigt lågt.



Figur 7:Roc-kurva för träningsdata

I figur 7 som visas ovan, syns Roc-kurvan för samma modell med justerade träningsdata och dess anpassade värden. Tröskelvärdet för denna del av datamaterialet är högre men fortfarande väldigt lågt (0,241).

5. Diskussion

I det femte kapitlet kommer en djupare tolkning av resultaten, granskning av resultatet och metoderna samt lämpligheterna av de använda metoderna att gås igenom.

För modellen som samtliga prediktioner och analyser baseras på syns ett positivt intercept för referensgruppen, vilket innebär att en passagerare som tillhör referensgruppen har en högre chans att överleva än att omkomma, alltså kvinnor. Även variabeln ålder har en positiv koefficient som innebär att ju äldre personen är desto högre sannolikhet att överleva. För övriga två variabler, män och interaktionen mellan män och ålder, minskar sannolikhet att överleva ju äldre männen blir.

Det som alltså framgår i resultatdelen är att kvinnornas överlevnadschans är hög och ökar ju äldre de blir medan männens överlevnadschans är låg och minskar ju äldre de blir. Detta framkommer av koefficienterna i **Tabell 2**: Modellen och oddskvoterna i Fel! Hittar inte referenskälla. där det syns en ökning av kvinnornas- och en minskning av männens överlevnadschans i samband med ålder. Det finns även stöd för detta i diagrammet i **Figur 3**.

Modellen har tränats på 70 % av datamaterialet och sedan testats på 30 % av datamaterialet, detta för att som nämnt i (**3. 1 Logistisk regression**) undvika överanpassning vad gäller modellen samt kunna utvärdera modellen och se hur bra den presterar utifrån utvärderingsmåtten för modellen. Uppdelning av datamaterialet leder delvis till att det blir svårt att använda sig av de låga tröskelvärden som returneras i (**4.2.3 ROC-kurvor**) eftersom datamaterialet är obalanserat, medan överanpassning undviks. Med hänsyn till ovanstående problematik har valet att ha 0,5 som tröskelvärde för observationernas sannolikhetsberäkning funnits lämpligt.

I samband med att huvudfokus låg på effekten av variablerna kön och ålder, leder detta till att variation i överlevnadsgrad som skulle kunnat förklarats av exkluderade variabler istället förklaras av enbart variablerna kön, ålder och dess interaktionstermer, ett bra exempel på detta är tabell 5 (**4.2.2 Sannolikhetsberäkningar**), där passagerare predikteras till att överleva baserat på dennes ålder och kön, detta eftersom att modellen som används enbart har två variabler därav tas inte andra faktum som passagerarnas reseklass eller hur mycket passagerarens resa kostade, i hänsyn. Alltså generaliseras utfallet utifrån att enbart ålder och kön används som förklarande variabler.

Vid valideringen av modellen skapad på testdata finner vi att prediktionerna har 0,6 procentenheters högre fel, modellen har lite svårare att prediktera rätt vad gäller överlevande passagerare och kan i 33,9 % av fallen prediktera en passagerare till att inte överleva trots att denne gjorde det. Trots detta presterar modellen bättre vad gäller att prediktera passagerare till att omkomma, vilket den gör med 85,6% säkerhet. Sammanfattningsvis är detta inte en särskilt tillförlitlig modell då det finns fler variabler som bör tas hänsyn till men exkluderats här då det inte ryms inom ramen för denna uppsats.

Modellen som baserats på data från Titanic visar hur det finns stora skillnader i överlevnadsgrad vad gäller ålder och kön där kvinnor och barn har en högre överlevnadsgrad, detta behandlar frågeställningarna kring huruvida det fanns skillnader mellan kön och åldersgrupper vad gäller överlevnadsgrad, ombord detta fartyg, där slutsatsen är att kvinnor

och barn har en bättre chans att överleva. Tidigare forskning behandlar samma frågeställning, vad gäller just Titanic, där en skillnad i överlevnadsmönster kan observeras, däremot är deras slutsats att detta gäller för just Titanic men ser tvärtom ut där kvinnor och barn har en sämre överlevnadsgrad annars (Erixson, 2012).

6. Slutsatser

Syftet med rapporten var att undersöka hur överlevnadschansen påverkas av kön och/eller ålder.

Kvinnor har i samtliga fall av analyserna haft en signifikant högre överlevnadsgrad. Alltså skiljer sig sannolikheten att överleva könen emellan, där kvinnor har en högre chans att överleva än omkomma till skillnad från män där det är tvärtom.

Interaktionen mellan kön och ålder har ett starkt samband med responsvariabeln. Detta då kön som är en starkt signifikant variabel tillsammans med ålder som inte är signifikant, kan skapa en signifikant påverkan på överlevnadschansen. Detta i form av att överlevnadschansen generellt sett är signifikant lägre för män. Interaktion hjälper även till att visa hur till exempel pojkar 0—15 år gamla har högre överlevnadschans än män äldre än dem.

Bibliografi

- ASELA, 2017. *Exploratory Data Analysis on the Titanic Dataset*. [Online]
  Available at: https://www.kaggle.com/code/aselad/exploratory-data-analysis-on-the-titanic-dataset/notebook
  [Accessed 30 Mars 2022].
- David G. Kleinbaum, M. K., 2010. *Logistic Regression: A Self-Learning Text*. 3 ed. s.l.:Springer-Verlag New York.
- Elinder, M. & Erixson, O., 2012. *Gender, Social Norms, and Survival in Maritime Disasters*. [Online]
  Available at:
  https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2161890_code685575.pdf?abstractid=2161890&mirid=1
  [Accessed 9 May 2022].
- Erixson, M. E. O., 2012. *Rädda sig den som kan! Om kön, normer och överlevnad vid fartygsolyckor,* s.l.: s.n.
- H.Kutner, M. C. J., 2005. *Applied Linear Statistical Models*. 5 ed. s.l.:s.n.
- Hietala, I., 2020. *Klassificering*. [Online]
  Available at: https://www.isakhietala.com/teaching/732g12/1-klassificering/
  [Accessed 26 Maj 2022].
- John Fox, S. W., 2018. *An R Companion to Applied Regression*. 3:e ed. s.l.:SAGE.
- Klienbaum, K. N. R., 2021. *Applied regression analysis and other multivariable methods*. 5th ed. Boston: Cengage.
- SBU, 2020. *SBU:s metodbok*. [Online]
  Available at: https://www.sbu.se/metodbok.
  [Accessed 5 Maj 2022].
- Stanford University, 2016. *A Titanic Probability*. [Online]
  Available at:
  https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/stuff/titanic.csv
  [Accessed 12 April 2022].
- Thoma, M., 2018. *WIKIPEDIA*. [Online]
  Available at:
  https://upload.wikimedia.org/wikipedia/commons/thumb/1/13/Roc_curve.svg/1024px-Roc_curve.svg.png
  [Accessed 11 Maj 2022].
- Ulfvarson, A., 2022. *Titanic*. [Online]
  Available at: http://www.ne.se/uppslagsverk/encyklopedi/lång/titanic
  [Accessed 25 Februari 2022].

Bilagor

Bilaga A

R-kod

```
#Paket

libraries <- c("cowplot", "tidyverse", "dplyr", "ggplot2", "knitr","matlib"
,"PerformanceAnalytics","car","corrplot","olsrr","leaps","rms","qpcR","MASS
","Caret","InformationValue","ISLR","bestglm","plyr","MASS","aod","survey",
"MuMIn","xfun","questionr")#Här tilldelar vi programmen som vi behöver för
koden.

#Databehandling
Titanic <-read.csv("https://web.stanford.edu/class/archive/cs/cs109/cs109.1
166/stuff/titanic.csv")
lapply(libraries, require, character.only = TRUE)
options(digits = 3)
Titanic$Age <-round(Titanic$Age,digits = 0)



# Figur 1

ggplot(Titanic, aes(x = Age, fill = as.factor(Survived))) +
  geom_histogram(position = "identity", alpha = 0.3,bins = 30)+xlab(label =
"Ålder")+
  ylab(label = "Antal")



levanede_man <- as.numeric(sum(Titanic$Survived == 1 & Titanic$Sex == "male
"))

levande_kvinna <- as.numeric(sum(Titanic$Survived == 1 & Titanic$Sex == "fe
male"))

död_man <- as.numeric(sum(Titanic$Survived == 0 & Titanic$Sex == "male"))

död_kvinna <- as.numeric(sum(Titanic$Survived == 0 & Titanic$Sex == "female
"))

l <-rbind.data.frame(levande_kvinna,levanede_man,död_kvinna,död_man)

u <-c(314,573,314,573)

s <-c("Kvinna","Man","Kvinna","Man")

y <-c("överlevde","överlevde","Omkom","Omkom")

c <-cbind.data.frame(l,s,y)

colnames(c) <-c("Antal","Kön","Lever")

u <-c(314,573,314,573)
```

```
c$Antal <-c$Antal/u

c$Kön <-as.factor(c$Kön)

c$Lever <-as.factor(c$Lever)




#Figur 2

ggplot(c, aes(fill=Lever, y=Antal, x=Kön)) +
  geom_bar( stat="identity")+
  labs(y="Procent")+
  scale_y_continuous(labels = scales::percent)+
  theme_bw()+ggtitle(label="Andel passagerare grupperat på överlevnadsgrad
och kön")


#Figur 3

ggplot(Titanic)+ geom_histogram(aes(x=Age, fill=Survived),

                    colour="grey50", alpha=0.3, position="identity")+ylab(la
bel="Antal")+theme_bw()+

    ggtitle(label="Antal passagerare grupperat på ålder, kön och överlevnad
sgrad")+xlab("Ålder")+labs(fill='Överlevnadsgrad')+facet_grid(~Sex)



#Logistisk regressionsmodellering


default <-Titanic

library(InformationValue)


#Delar upp datasetet I tränings och testdata för att kunna utvärdera modell
en


data <-default


#Delar upp datasetet I tränings och testdata för att kunna utvärdera modell
en

set.seed(3246462) #Sätter seed


sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE,prob=c(.7,.3)) #S
amplar observationer där 70 % är träningsdata och 30 % är testdata
train <- data[sample, ] #Delar upp datasetet I tränings och testdata för at
```

```
t kunna utvärdera modellen.
test <- data[!sample, ]


# Tabell 2
model1 <-glm(Survived~Age*Sex,family=binomial(link = "logit"),data=train)



#Skapar prediktionerna som jämförs mot de sanna utfallen
pred_y <- predict.glm(model1, newdata=test)


#Konvertar de anpassade värdena till 1:or och 0:or

Model1anpassad <- ifelse(model1$fitted.values > 0.5, 1, 0)



#Konvertar defaults från "Yes" and "No" till 1:or och 0:or
test$Survived <- ifelse(test$Survived > 0.5, 1, 0)


Tabell 5
confusionMatrix(test$Survived, pred_y)


Tabell 6
confusionMatrix(train$Survived,model1anpassad)

#Missklassificeringsgrad för modelllen (Först test, sen träning)
misClassError(test$Survived, pred_y, threshold=0.5)

misClassError(train$Survived,model1anpassad, threshold=0.5)

#Känslighet för modelllen
sensitivity(test$Survived, pred_y, threshold=0.5)
sensitivity(train$Survived,model1anpassad, threshold=0.5)

#Specifitivitet för modelllen
specificity(test$Survived, pred_y, threshold=0.5)
specificity(train$Survived,model1anpassad, threshold=0.5)

#Figur 6

library(pROC)

plot.roc(test$Survived, pred_y, percent = FALSE, main = "ROC curves", add =
FALSE, asp = NA, print.auc = TRUE)

legend("bottomright",

       legend = c("Testdata"),

       col = c("black"),

       lty = c(1))
```

```r
#Figur 7

plot.roc(train$Survived, model1anpassad, percent = FALSE, main = "ROC curve
s", add =  FALSE, asp = NA, print.auc = TRUE)

legend("bottomright",

       legend = c("Träningsdata"),

       col = c("black"),

       lty = c(1))


#Tabell 4
r <-model1$coefficients
t <-odds.ratio(model1)[,1]
k <-cbind(r,t)
colnames(k) <-c("Koefficient","Oddskvot")


#Tabell 7

set.seed(3249465)
test$probpred <-round((pred_y)3)
test$predikt<- ifelse(test$probpred > 0.5, 1, 0)
TABELL6 <-sample_n(test,6,replace = FALSE)
data.frame(TABELL6)


#Tabell 3

vif <-car::vif(model1)
data.frame(vif)
```