UPTEC STS 09036
Examensarbete 30 hp
Oktober 2009

# Statistisk analys och modellering av kreditrisker

Pontus Ahlqvist

UPPSALA
UNIVERSITET

**Teknisk- naturvetenskaplig fakultet**
**UTH-enheten**

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

# Abstract

## Statistisk analys och modellering av kreditrisker

## Statistical analysis and modelling of credit risks

*Pontus Ahlqvist*

This master thesis project in mathematics focuses on statistical modelling and analysis of credit risks. The project was carried out in collaboration with a Nordic financial corporation which in its daily business operation offers short consumer credits to its customers. The company is thus heavily exposed to credit risks which need to be managed in an effective way.

The paper investigates how classification algorithms (using logistic regression) can be implemented and used in order to distinguish "good" (Group 1) from "bad" (Group 2) customers. In addition, the paper also includes quantitative analysis as well as identification of risk factors (using cluster analysis), within the group of "bad" customers.

The analysis and modelling are based on real-life data which were extracted from the company's databases. The data consists of about 1300 observations from the two groups and nine different variables were available for each observation.

The main finding of the research is that the two groups of observations are somewhat similar and it is therefore fairly difficult to construct good classification models. The best model developed, classifies correctly in about 7 cases out of 10.

An insight from the cluster analysis is the difficultly of identifying clear clusters with the number of variables and observations that were available. In order to construct accurate clusters, more observations (from the group consisting of "bad" customers) and variables must be included in the analysis. However, some interesting variable combinations were identified using the correlation- and cluster analysis. One of them is the identification of the potential risk group consisting of old people applying for credits in the end of the months. Another interesting insight is the fact that the variables age, score and income were heavily correlated for the "bad" customers which was not the case for the other group.

Handledare: Sekretessbelagt
Ämnesgranskare: Jesper Rydén
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 09036

# Statistisk analys och modellering av kreditrisker

Pontus Ahlqvist

7 oktober 2009

**Sammanfattning**

Detta examensarbete avhandlar hur man genom att studera mönster hos historiska betalningstransaktioner kan bedöma hur troligt det är att en ny kund återbetalar sina skulder. Detta avgörs genom att studera kundspecifika variabler (information om kunden som finns tillgänglig).

Examensarbetet genomförs vid ett finansbolag i Stockholm vilket bedriver verksamhet som är förenad med att låna ut pengar. Företaget har därmed stor exponering mot risken att återbetalning av lånet inte sker (så kallad kreditrisk). Om företaget beviljar kredit (lånar ut pengar) till en person som med stor sannolikhet inte kommer att betala tillbaka lånet, kan det innebära förluster för företaget. Därför genomför företaget en kreditprövning av varje ny kund innan kredit beviljas. Denna kreditprövning genomförs av ett externt kreditupplysningsföretag som poängsätter kundens troliga återbetalningsförmåga (där 100 poäng är bäst och 0 är sämst) . Denna poäng ligger sedan till grund för att avgöra hur stor kredit kunden kan erhålla.

Tidigare studier har visat att denna poäng troligen kan kombineras med ytterligare information om kunden, för att på så sätt ännu bättre kunna avgöra hur troligt det är att kunden kommer att återbetala sina skulder. Syftet med examensarbetet är att undersöka hur denna ytterligare information, bestående av 8 variabler i kombination med kundens erhållna poäng, kan användas för att göra bedömningen av trolig återbetalning ännu mer träffsäker.

Resultaten av examensarbetet visar hur modeller för beräkning av trolig återbetalningsförmåga blir bättre då kombinationer av dessa ytterligare variabler används. Negativt är att dessa modeller beräknar fel i ungefär 3 fall av 10. En av slutsatserna blir därför att det är svårt att bygga allmängiltiga modeller varför det istället är bättre att fokusera på att kombinera poäng med enstaka delintervall av variabelkombinationer för att på så sätt identifiera intressanta samband. Examensarbetet identifierar också några sådana kritiska variabelkombinationer, vilka kan användas i kreditgivningsprocessen för att avgöra vem som kan erhålla kredit, och i så fall, storleken på denna kredit.

# Innehåll

# Tabeller

# Figurer

# Kapitel 1

# Introduktion

## 1.1 Bakgrund

Kreditrisk är en av de många risker finansbolag måste hantera vid utlåning till såväl företag som privatpersoner. Risken definieras som risken att motparten i en transaktion inte fullgör sina förpliktelser [12]. Om låntagaren inte lyckas med dessa förpliktelser anses låntagaren ha falsifierat (eng. *default*). Företaget som detta examensarbete genomförs i samarbete med tillhandahålla korta konsumentkrediter för sina slutkunder och denna affärsverksamhet innebär därmed stor exponering mot kreditrisker.

Innan fallissemang finns det inget otvetydigt sätt att skilja på låntagare som kommer fullfölja sina åtaganden och de som inte kommer att göra det. Det bästa som kan göras är att använda teoretiska sannolikhetsmodeller (baserat på historiska erfarenheter) för att på så sätt avgöra hur troligt det är att en ny låntagare kommer att fullfölja åtaganden eller inte. En låntagare betalar därför ofta en premie (utöver den riskfria räntan) som är proportionell till dennes sannolikhet för fallissemang för att på så sätt kompensera kreditorn för dess risker med utlåningen. [4]

Då företaget hanterar ett stort antal transaktioner tillsammans med det faktum att beloppen som lånas ut får anses relativt små baseras premien inte på varje specifik låntagares troliga betalningsförmåga, istället betalar alla kunder en och samma premie vilket är inbakat i de avgifter företaget tar för att hantera varje enskild transaktion.

Examensarbetet handlar om hur dessa ovan diskuterade teoretiska sannolikhetsmodeller kan se ut och hur analys av historiska transaktioner kan användas för att förbättra de modeller som används hos företaget idag. I dagsläget används primärt den poäng som varje kund får vid en kreditprövning för att bedöma en ny kunds troliga återbetalningsförmåga. Det finns dock tidigare studier (se avsnitt 1.3) som indikerar att denna poäng kan kombineras med annan kundspecifik information för att på så sätt göra uppskattningar ännu mer träffsäkra.

## 1.2 Syfte

Syftet med uppsatsen är tvådelat varav det första syftet är att analysera två grupper av kunder hos företaget. Den ena gruppen (Grupp 1) består av kunder som betalar sina skulder i tid och den andra gruppen består av de kunder som inte betalat sina skulder överhuvudtaget (Grupp 2), se figur 2.1. Genom att jämföra dessa grupper och identifiera skillnader och likheter (utifrån historiska observationer av kundspecifika variabler) är förhoppning att ge företaget djupare insikter i deras slutkunders köp och betalningsbeteende. Inom respektive grupp genomförs också analyser av korrelationer mellan studerade variabler för att i bättre utsträckning förstå hur kundspecifika variabler är relaterade med varandra. Analysmetoder för att undersöka detta innefattar bland annat hypotesprövningar, regressionsanalyser och klusteranalyser.

Det andra syftet är att undersöka om det rent matematiskt går att konstruera en allmängiltig modell för hur nya kunder kan klassificeras in i någon av dessa två grupper. Genom att använda tillgänglig information om historiska transaktioner är förhoppningen att kunna ge förslag på hur nya klassificeringsmodeller skulle kunna konstrueras.

## 1.3 Tidigare studier

Mycket i denna studie bygger vidare på studien *Consumption credit default predictions* [8] där författarna konstaterar hur trolig återbetalningsförmåga bedöms bättre om icke-traditionella variabler såsom exempelvis tidpunkt för beställning, geografisk bostadsort eller en kunds e-postadress används vid bedömningen. Dessa variabler ingår traditionellt inte i den poäng som kreditupplysningsföretag tillhandahåller. Totalt används 27 olika variabler.

Författarna konstruerar i studien tre olika regressionsmodeller där de konstaterar hur den modell som innehåller dessa icke-traditionella variabler ger en bättre passningsgrad än de övriga två modellerna. Exempelvis konstateras i studien hur yngre personer är sämre att betala sina skulder såväl som hur personer bosatta på landsbygden är bättre på att betala sina skulder än de som bor i storstäder.

Vad som primärt skiljer detta examensarbete från den tidigare studien är att ta steget från att konstatera att en variabel är signifikant till att inkludera dessa slutsatser i en klassificeringsmodell för klassificering av nya kunder. En ytterligare vidareutveckling är de klusteranalyser som genomförs i syfte att försöka identifiera extra kritiska och homogena kundgrupper där betalningsförmågan (eller betalningsviljan) är extra dålig.

# Kapitel 2

# Metod

## 2.1 Angreppssätt och disposition

Strukturellt är studien uppdelad i tre faser där de första två faserna är relaterade till det första syftet (analys av de två grupperna) och den tredje fasen är relaterad till uppsatsens andra syfte (konstruering av klassificeringsmodell). Faserna genomförs i sekvens då den tredje fasen bygger på erfarenheter och slutsatser från de två tidigare faserna. De tre faserna är som följer:

1. Att genom signifikanstester undersöka om det föreligger någon signifikant skillnad i fördelning mellan observationer för respektive grupp. Detta upprepas och genomförs parvis för respektive variabel. Dessa signifikanstester kompletteras också med hypotesprövningar (vilka bygger på logistisk regressionsanalys) för att på så sätt undersöka vilka variabler som föreligger vara signifikanta vid bedömning av kreditvärdighet. Denna fas påminner mycket om de analyser som genomfördes i [8] med skillnaden att testerna genomförs med en annan analysmetod och med en annan geografisk avgränsning.

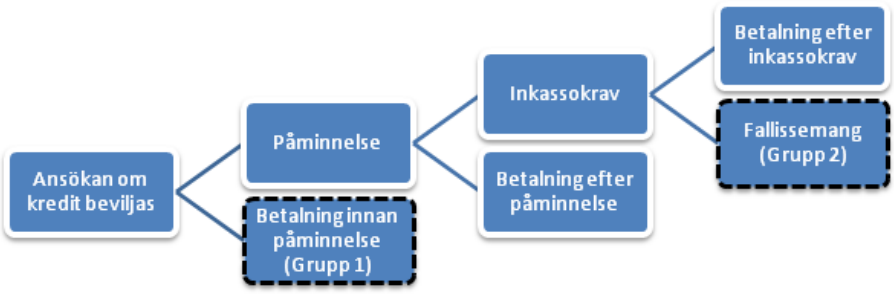
2. Att genom korrelations- och klusteranalyser undersöka, hur de i studien ingående variablerna och observationerna, är korrelerade med varandra. Dessa två analysmetoder kompleterar varandra och förhoppningen är att med hjälp av dessa analyser kunna identifiera intressanta klustertendenser hos observationer i grupp 2. Dessa insikter kan sedan användas av företaget för att avgöra om en kund ska tillåtas kredit eller inte.

3. Att genom analys av historiska transaktioner utveckla en modell vars syfte är att kunna klassificera nya kunder till någon av de två grupperna. Genom att kombinera den poäng varje kund erhåller med andra kundspecifika variabler är förhoppningen att denna utbyggda modell, med så hög träffsäkerhet som möjligt, ska kunna användas för att kunna klassificera nya kunder till någon av de två grupperna.

## 2.2 Definition av analyserade grupper

En förenkling som gjorts i denna uppsats är att endast använda två grupper vid analys och klassificering av observationer. I verkligheten existerar dock fler möjliga grupper och i figur 2.1 beskrivs ett exempel på hur kredits levnadsprocess kan se ut. I figuren framgår tydligt hur det finns fyra slutgiltiga tänkbara grupper och det är två av dessa (Grupp 1 och Grupp 2 ) denna studie fokuserar på. Att klassificera till fler än två grupper är möjligt men då antalet observationer är begränsat har denna avgränsning gjorts i samråd med handledare och ämnesgranskare. I tabell 2.1 finns en sammanställning över hur många observationer som återfinns i respektive slutgiltig grupp. De två grupperna som fokus ligger på i denna studie definieras enligt:

- **Grupp 1** är krediter som har återbetalats innan första påminnelse har gått ut till kund. Till denna grupp hör de kunderna som är allra duktigast på att betala och ingen hänsyn tas till hur lönsamma dessa kunder är (exempelvis storlek på kredit, hur tidigt eller sent under återbetalningsperioden krediten återbetalas etcetera). Dessa krediter utgör 88 procent (1016 till antalet) av det totala antalet observationer.
- **Grupp 2** är krediter som inte har betalts inom 60 dagar efter att krediten har beviljats. Detta innebär också att krediten gått från inkasso till att blivit såld på andrahandmarknad för obetalda fakturor. Dessa observationer utgör 12 procent (140 till antalet) av det totala antalet observationer. Inte heller här görs någon bedömning av hur lönsamhet hos dessa krediter.




Figur 2.1: En kredits olika stadier

| Grupp | Antal observationer |
|---|---|
| Ansökan om kredit beviljas | 1693 |
| Betalning innan påminnelse (Grupp 1) | 1016 |
| Betalning efter påminnelse | 500 |
| Betalning efter inkassokrav | 37 |
| Fallissemang (Grupp 2) | 140 |

Tabell 2.1: Antal observationer för respektive slutgrupp hos krediter

## 2.3 Variabler som ligger till grund för analyser

Eftersom data är obalanserat (i den bemärkningen att de båda grupperna är olika stora med avseende på antal observationer, se tabell 2.1) krävs lite extra funderingar kring hur många variabler som bör ingå i analyser och modeller. Är observationerna för få och variablerna för många finns en risk för systematiska fel (eng. *bias*) i ML-skattningar (eng. *maximum likelihood estimates*) av $\beta$-koefficienterna (se Agresti [1] för diskussion om systematiska fel samt avsnitt 2.5.3 för diskussion om ML-skattningar och $\beta$-koefficienter). Agresti skriver hur antalet variabler approximativt bör motsvara ungefär den procentsats den mindre gruppen utgör av hela populationen. Detta innebär att för det dataset som används i denna uppsats bör ungefär tolv variabler (grupp 2 utgör 12 procent av hela populationen, $140/(140+1016) \approx 12\%$) användas för att på så sätt minska risken för systematiska fel.

Initialt var tanken att analyser i denna uppsats skulle baseras på fler än de åtta variabler som presenteras nedan, men på grund av problem med att extrahera data från servrar blev undersökningen begränsad till dessa variabler. Nedan presenteras de ingående variablerna, där $X_7$ , $X_8$ är kategorivariabler och övriga är mätvariabler.

- $X_1 = Datum$ Variabeln anger vilken veckodag kunden har ansökt om krediten. Variabeln antar reella heltalsvärden i intervallet $[1-31]$.
- $X_2 = Tidpunkt$ Variabeln anger vilken tidpunkt kunden har ansökt om krediten. Variabeln antar reella heltalsvärden i intervallet $[0-23]$.
- $X_3 = Belopp$ Variabeln anger storleken på krediten som kund erhållit. Variabeln antar reella heltalsvärden enligt $[179, 249, 699, 477, 995]$. (Variabeln kodas med automatik som 1,2,3,4 eller 5 i statistikprogrammet R)
- $X_4 = Ålder$ Variabeln anger ålder på kund som beviljats krediten. Variabeln antar reella heltalsvärden $\geqslant 18$.
- $X_5 = Poäng$ Variabeln anger den poäng som varje kund erhåller vid kreditprövning via externt kreditupplysningsföretag. Variabeln antar reella heltalsvärden i intervallet $[0-100]$, där 100 är bäst och 0 är sämst. Poängen indikerar på kundens troliga återbetalningsförmåga enligt bedömning gjord av kreditupplysningsföretaget. Exakt vad denna poäng baseras

på och hur modellen är viktad är dock kreditupplysningsföretagets affärshemlighet. Troligt är dock hur variabler såsom exempelvis ålder och inkomst har betydelse för poängsättning (se avsnitt 2.5.2.1 för diskussion om multikollinearitet).

- $X_6 = Inkomst$ Variabeln anger kundens årsinkomst vilket kan anta alla reella heltalsvärden $\geqslant 0$.
- $X_7 = NyKund$ Variabeln anger om kunden ansökt om kredit tidigare. (Variabeln kodas med automatik som 0 eller 1 i statistikprogrammet R)
- $X_8 = Kön$ Variabeln anger kön hos kund. (Variabeln kodas med automatik som 0 eller 1 i statistikprogrammet R)

## 2.4 Motiv för val av data

Datamängden består totalt av 1156 transaktioner varav 1016 observationer tillhör grupp 1 och 140 observationer tillhör grupp 2 (se tabell 2.1). Den första december 2008 genomförde företaget förändringar i de kriterier som avgör vilka kunder som beviljas krediter. Dessa ändringar gav ett positivt utfall och att blanda ihop data från en period innan dessa förändringar med data efter förändringarna skulle innebära onödiga risker för feltolkningar. Därför analyseras endast krediter som beviljats efter den 15 december 2008.

Det finns alltid en viss fördröjning från det att en kredit beviljats till att den kan anses ha falsifierat. Vilket åskådliggörs i figur 2.1 utgår det påminnelser om betalning vid flera tillfällen. Efter noggranna diskussioner med framförallt handledare[1] anses en kredit ha falsifierat om den inte betalats inom 60 dagar efter att den skapats och kan därmed anses tillhöra grupp 2. Slutgiltigt dataset som använts för analyser extraherades den 15 maj 2009 och innehåller därmed krediter som skapats mellan den 15 december 2008 och 15 mars 2009.

En kredit som är obetald och där det går ut påminnelser behöver inte nödvändigtvis vara en dålig affär för företaget. Påminnelser och straffavgifter för obetalda skulder kan ibland vara gynnsamma affärer då beloppen för krediter är förhållandevis små och avgifterna för förseningar och påminnelser är förhållandevis höga. När en kredit blir obetald en längre period (60 dagar och längre) är dock sannolikheten för återbetalning av krediten väldigt liten och innebär därför förlust för företaget[2].

Analys kring lönsamhet hos krediter som befinner sig mellan de båda grupperna (Grupp 1 och Grupp 2) kan göras men är inte aktuellt för detta examensarbete. Avgränsning har alltså gjorts utifrån de mest extrema grupperna (de som är allra bäst respektive allra sämst på att betala).

[1] Utifrån diskussion med handledare 2009-04-15
[2] Ibid

## 2.5 Motiv för val av teoretiskt angreppssätt

I avsnitt 2.5.1-2.5.3 resoneras kring vilka statistiska metoder och modeller som anses vara användbara för de undersökningarna som krävs för att besvara syfte och frågeställning. De statistiska testerna och matematisk modellering har gjorts i statistikprogrammet R (version 2.6.2.)[3]. För källkod till de program som skrivits för examensarbetet, se bilaga A. Samtliga tester genomförs på 95 procent säkerhetsnivå.

### 2.5.1 Signifikanstester och hypotesprövningar

Förhoppningen med dessa tester är att nya hitintills okända variabler framträder som signifikanta och därmed kan inkluderas i de modeller som avgör vem som beviljas kredit. Signifikanstesterna kompletteras med hypotesprövningar (vilka undersöker eventuella skillnader i fördelning hos variablerna mellan de båda grupperna) och utgår från följande hypoteser:

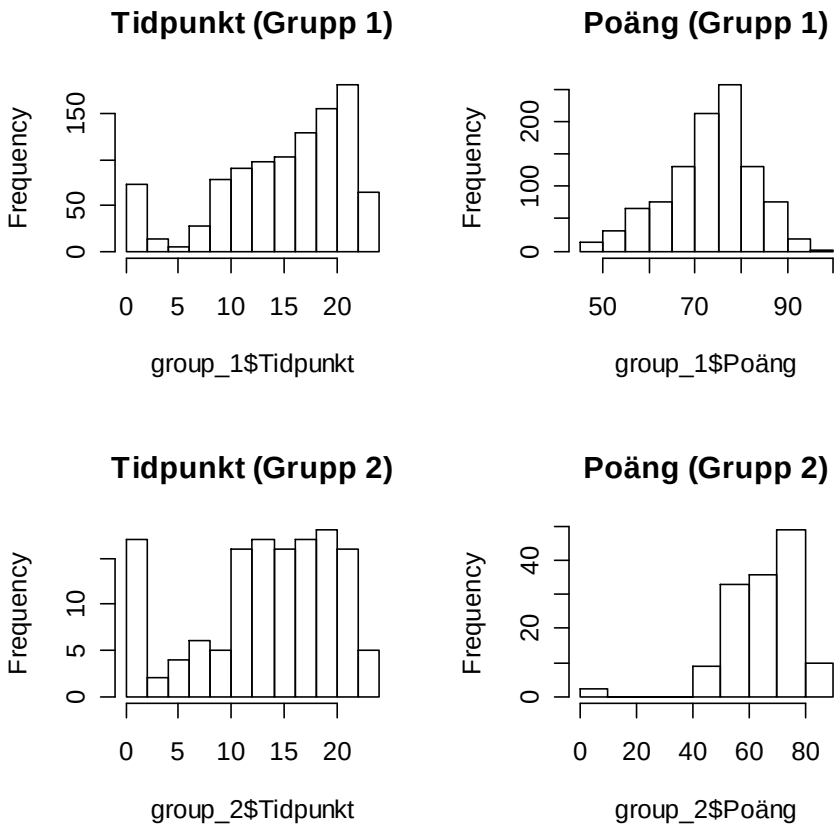
- $H_0$ : *Det föreligger ingen skillnad i fördelning mellan de båda grupperna*
- $H_1$ : *Det föreligger en skillnad i fördelning mellan de båda grupperna*

För att avgöra vilket statistiskt test som är lämpligt för att genomföra hypotesprövningarna studeras först täthetsfunktioner för respektive variabel och grupp. Exempel på hur dessa täthetsfunktioner kan se ut för två av variablerna för respektive grupp finns i figur 2.2.

För att underlätta jämförelser mellan olika variabler, är det önskvärt att genomföra samma statistiska test för samtliga kontinuerliga variabler. För detta krävs dock att variablerna antar samma fördelning, vilket inte är fallet (se figur 2.2). Ett icke-parametriskt (rangsummetest) test är därför det enda alternativet för de kontinuerliga variablerna. Det finns olika varianter men eftersom data utgörs av två oberoende populationer (grupper) måste testet beakta detta. Ett lämpligt alternativ blir därför Wilcoxons tvåsticksprovstest (även kallat Mann-Whitneys test) [13]. De två kategoriska variablerna analyseras med hjälp av chi-2 test.

Vid genomförandet av dessa två tester returneras i R ett så kallat p-värde vilket jämförs med vald säkerhetsnivå (95 procent i detta fall). Nollhypotesen $H_0$ förkastas om p-värdet $< 0.05$.

[3]Finns att ladda ner gratis på `http://www.r-project.org`




Figur 2.2: Exempel på täthetsfunktioner för två variabler ($X_2$ och $X_5$) för respektive grupp

### 2.5.2 Korrelations- och klusteranalyser

#### 2.5.2.1 Korrelationsanalyser

Till skillnad från signifikanstester och hypotesprövningar (där de båda grupperna jämförs mot varandra) genomförs i denna fas analyser av korrelationer inom respektive grupp . Är det exempelvis så att ålder och när på dygnet en kreditansökan inkommit är högt korrelerade för grupp 2 men inte för grupp 1, är denna insikt viktig då företaget kan utveckla så kallade *reject-kriterier* (kombinationer av variabler som ovillkorligen nekas krediter) utifrån dessa insikter. Vid analys av kontinuerliga variabler studeras korrelationsmatriser och för kategoriska variabler används korstabeller. För rent matematiska definitioner av hur korrelationskoefficienter beräknas, se avsnitt 3.3.1-3.3.2.

Eftersom det inte är känt vad variabeln $X_5$ (Poäng) baseras på finns en in-

byggd risk för multikollinearitet. Multikollinearitet kan inträffa när två eller flera variabler, som ingår i en regressionsmodell, är starkt korrelerade med varandra. Multikollinearitet är alltså risken för att två variabler till viss del innehåller samma information [11]. Ett sätt att identifiera multikollinearitet är att vid korrelationsanalyser vara uppmärksam på höga korrelationskoefficienter, detta är dock ingen garanti för att multikollinearitet inte föreligger. Det är svårt att ange någon exakt gräns för vid vilka värden på korrelationskoefficienten risk finns för multikollinearitet med en tummregel kan vara att iakttaga försiktighet då $\rho > 0.75$ [5].

#### 2.5.2.2 Klusteranalyser

Då korrelationsmatrisen endast beaktar två variabler åt gången blir dessa analyser något begränsade då det uppenbarligen kan finnas multivariata samband av intresse. Dessa samband kan studeras med klusteranalys vars syfte är att gruppera kunder (objekt) med liknande egenskaper tillsammans [9].

Genom att jämföra initiala korrelationer (för grupp 2) med de korrelationer som uppträder i respektive kluster för samma grupp (se avsnitt 2.5.2.2) kan insikten i vilka variabler som utgör respektive kluster förbättras. Rimligtvis bör respektive kluster innehålla högre korrelationer än gruppen som helhet eftersom liknande objekt sorterats in i respektive kluster utifrån deras egenskaper. Objekt i samma kluster har likheter med varandra och dessa likhet mäts utifrån deras korrelation med varandra.

För att konstruera kluster finns olika metoder att tillgå (exempelvis hierarkiska kluster eller K-mean kluster) och i R finns inbyggda funktioner för dessa analyser. För att utveckla klusteralgoritmer krävs först och främst en definiering för hur de många objekten ska grupperas; utifrån likheter (eng. *single*), olikheter (eng. *complete*), ett medel av dessa två (eng. *average*) eller förlustförminskande (eng. *Ward*). För att mäta hur lika (eller olika) dessa objekt är, definieras också deras inbördes avstånd med lämplig avståndsmetodik (exempelvis euklidiskt-, Minkowski- eller manhattanavstånd). [7]

Fördelen med hierarkiska kluster är att denna metod inte kräver några fördefinierade grupper vilket K-mean metoden gör [7]. I [9] konstateras också hur euklidiska avstånd är det bästa avståndsmåttet vid explorativa klusteranalyser (se avsnitt 3.3.3 för matematisk definition).

Eftersom undersökningarna i denna uppsats är väldigt explorativa till naturen (inga initiala antaganden har gjorts beträffande kluster eller fördelningar) kommer hierarkiska kluster användas där avstånden mellan objekt mäts med euklidiska avstånd.

Den hierarkiska kluster metodiken kan visualiseras i ett så kallat dendrogram där objekten delas upp enligt en omvänd trädstruktur där varje vertikal nivå representerar en speciell nivå av likhet (y-axel) . Desto närmare objekt befinner sig horisontellt desto närmare befinner de sig varandra (x-axel). Se figur 4.1 för exempel.

### 2.5.3 Klassificeringar med regressionsanalys

En klassificeringsmodell har som syfte att utifrån en uppsättning egenskaper hos en kund (där variabler antar olika värden), med så hög säkerhet som möjligt, avgöra vilken grupp kunden är mest trolig att tillhöra [9]. Att kunna klassificera nya kunder med 100 procents säkerhet vore idealt, men tyvärr ser verkligheten inte ut så och det finns alltid en viss risk för felklassificeringar. Genom att studera historiska transaktioner ökar dock kundkännedomen och möjligheterna till utveckling av bra klassificeringsmodeller förbättras.

Utgångspunkten är att använda historiska transaktioner (eng. *trainingset*) för bygga upp och kalibrera modellen. Denna modell testas sedan på ett antal observationer (eng. *testingset*) för att undersöka modellens träffsäkerhet. Det vanligaste alternativet är att använda någon form av regressionsmodell. En multipel linjär regressionsmodell kan byggas ut oändligt och skrivs på formen

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n. \tag{2.1}$$

I ekvation 2.1 motsvaras $X_i$ av de olika vektorerna med observationer för respektive variabel och $\beta$ är koefficienterna till dessa vektorer (alltså hur stor inverkan de har på slutgiltiga resultaten). $\beta_2$ avgör alltså hur stor inverkan $X_2$ har på responsvariabeln $Y$.

Då ingående variabler är både kontinuerliga och kategoriska krävs att den modell som används kan hantera denna blandning av skalnivåer (se avsnitt 3.1).

I fallet då responsvariabeln $Y$ är binär finns i princip bara en metod för att hantera denna blandning av skalnivåer och denna metod är logistisk regression. Flera tidigare studier har också konstaterat hur just logistisk regressionsanalys är den mest adekvata modellen för beräkning av kreditvärdighet [9, 1]. Därför kommer multipel logistisk regressionsanalys användas vid konstruering av klassifikationsmodeller.

# Kapitel 3

# Teori

## 3.1 Skalnivåer

Fyra olika skalor kan användas för att definiera den inbördes ordningen mellan de talvärden som en variabel kan anta [5]. I tabell 3.1 redogörs tre begrepp som används för att särskilja dessa skalnivåer.

Tabell 3.1: Mätskalor for variabler

| | Nominalskala | Ordinalskala | Intervallskala | Kvotskala |
|---|---|---|---|---|
| Rangordning | | X | X | X |
| Ekvidistans | | | X | X |
| Absolut nollpunkt | | | | X |

Begreppet kategorisk variabel används också, och syftar då på olika kategorier en variabel kan anta (exempelvis gul, grön, blå). Variabeln mäts alltid på nominalskala [5]. I detta arbete mäts variabel $X_7$ (NyKund) och $X_8$ (Kön) nominalskala och övriga variabler ($X_1 - X_6$) mäts på kvotskala. I vissa fall används också begreppet binära variabler vilket är ett smalare begrepp av kategoriska variabler då de endast kan anta två värden (exempelvis man eller kvinna). Dessa mäts också på nominalskala.

## 3.2 Signifikanstester och hypotesprövningar

### 3.2.1 Multivariat logistisk regression

I detta avsnitt presenteras logistisk regression närmare och formuleringarna är baserade på [9]. Istället för att modellera sannolikheten $p$ (för att en händelse inträffar) direkt med en linjär modell, betraktas först den så kallade oddskvoten

$$odds = \frac{p}{1-p} \tag{3.1}$$

där $p$ är sannolikheten för att en händelse inträffar. Kvoten består alltså av relationen mellan att en händelse inträffar mot att den inte inträffar. Ponera exempelvis att sannolikheten (baserat på kundens egenskaper och utifrån erfarenheter från historiska transaktioner) för att en kund tillhör grupp 1 är 80 procent och sannolikheten att en observation tillhöra grupp 2 är 20 procent. Detta skulle leda till oddskvot enligt

$$odds = \frac{0.8}{(1-0.8)} = 4. \tag{3.2}$$

I logistisk regression för binära responsvariabler ($Y$ antar antingen 1 eller 0), vilket är fallet med de två grupperna som studeras i detta arbete, modelleras den naturliga logaritmen av oddskvoten

$$logit\,(p) = ln\,(odds) = ln\left(\frac{p}{1-p}\right). \tag{3.3}$$

Logitfunktionen är en funktion av sannolikheten $p$. I den enklaste modellen antas hur *logit* visualiseras som en rak linje (med prediktorn $z$) enligt

$$logit(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 z. \tag{3.4}$$

Det är dock oftast lättare att tänka i termer av sannolikheter så genom omarbetning av ekvation 3.4 ges uttrycket för sannolikheten $p$ enligt

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z_1)}{1 + \exp(\beta_0 + \beta_1 z)}. \tag{3.5}$$

Genom att ansätta en logistisk regressionsmodell kan denna därmed arbetas om för att få fram en förväntad procentuell sannolikhet på hur troligt det är att något inträffar. I fallet med denna uppsats är denna sannolikhet kopplad till grupper och modellen kan användas för att beräkna hur sannolikt det är att en observation hamnar i grupp 1 respektive grupp 2 (se exempelvis avsnitt 3.4).

### 3.2.2 Hypotesprövningar för kontinuerliga variabler

För att undersöka om två stokastiska variabler har samma fördelning föreslår [2] följande tillvägagångssätt.

Antag att vi vill jämföra två stickprov, A: $x_1, \ldots, x_m$ från X, och B: $y_1, \ldots, y_m$ från Y, som omfattar $m$ respektive $n$ observationer och testa hypotesen $H_0$ : "$X$ och $Y$ har samma fördelning". Rangordna samtliga $N = m+n$ observationer och beräkna rangsumman, $R_A$ för $x$-stickprovet och $R_B = N\,(N+1)\,/2 - R_A$ för $y$-stickprovet. Normalt används den rangsumma som hör till det mindre av stickproven. För större stickprov utnyttjas att $R_A$ är approximativt normalfördelad,

$$R_A \approx N\left(\frac{m(N+1)}{2}, \frac{mn(N+1)}{12}\right). \tag{3.6}$$

Detta kan sedan göras om till standardiserad normalfördelning och $p$-värde slås utifrån detta upp i tabell. Genom att jämföra deta $p$-värde med vald säkerhetsnivå kan nollhypotesen $H_0$ antingen förkastas eller accepteras. Dessa beräkningar sker vanligtvis i statistiska program såsom exempelvis R.

## 3.3 Korrelations- och klusteranalyser

För att beräkna hur två datamängder samvarierar används korrelationskoefficienten $\rho$ som ett mått på samvariation. För korrelationskoefficienten $\rho$ gäller $-1 \leq \rho \leq 1$ där 1 anger perfekt positiv korrelation, -1 anger perfekt negativ korrelation och 0 anger att ingen korrelation existerar. I tabell 3.2 finns benämningar för respektive korrelationsintervall [11].

Tabell 3.2: Intervall för korrelationskoefficienter

| Korrelation | Negativ | Positiv |
|---|---|---|
| Svag korrelation | -0.3 till -0.1 | 0.1 till 0.3 |
| Viss korrelation | -0.5 till -0.3 | 0.3 till 0.5 |
| Stark korrelation | -1.0 till -0.5 | 0.5 till 1.0 |

### 3.3.1 Korrelation för kontinuerliga variabler

Om man kan anta att observationer är stickprov från de stokastiska variablerna $X$ och $Y$, kan man med hjälp av $\rho_{X,Y}$ testa om $X, Y$ är oberoende. Låt $((x_1, y_1), \ldots, (x_n, y_n))$ vara ett tvådimensionellt datamaterial omfattande $n$ observationer. Korrelationskoefficienten (Pearsons produktkorrelationskoefficient), eller kortare korrelationen, definieras då som

$$\rho_{X,Y} = \frac{cov\,(X,Y)}{\sigma_X \sigma_Y}. \tag{3.7}$$

Som skattning av $\rho$ används stickprovets korrelationskoefficient $r$,

$$r = S_{xy}/\sqrt{S_{xx}S_{yy}}$$

där $S_{xx}$ och $S_{yy}$ ges av ekvation 3.8 och 3.9

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.8}$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \tag{3.9}$$

Dessa korrelationskoefficienter sammanställs lämpligtvis i en matris enligt 3.10. I matrisen symboliserar exempelvis $r_{1,2} = r_{2,1}$ korrelationskoefficienten mellan vektor 1 och vektor 2 för de stokastiska variablerna $X$ och $Y$.

$$\begin{bmatrix} r_{1,1} & r_{1,2} & r_{1,3} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & & & \\ r_{3,1} & & \ddots & & \\ \vdots & & & \ddots & \\ r_{n,1} & & & & r_{n,n} \end{bmatrix} \tag{3.10}$$

#### 3.3.1.1 Test av likhet hos två korrelationskoefficienter

Antag att två oberoende stokastiska variabler $X$ och $Y$ utgörs av de parvisa observationerna $n_1$ och $n_2$ från en bivariat normalfördelad datamängd. Med hjälp av Fishers $Z$-transform kan nollhypotesen $H_0$ ($\rho_1 = \rho_2$) testas [3]. Fishers $Z$-transform ges utifrån ekvation 3.11-3.12, där $r_1$ respektive $n_1$ är den skattade korrelationskoefficienten respektive antalet observationer för grupp 1 (på samma sätt som $r_2$ och $n_2$ representerar liknande för grupp 2). $Z$ skattas då som

$$Z_1 = \frac{1}{2}\log_e\left(\frac{1+r_1}{1-r_1}\right) \tag{3.11}$$

$$Z_2 = \frac{1}{2}\log_e\left(\frac{1+r_2}{1-r_2}\right) \tag{3.12}$$

där $Z_1$ och $Z_2$ är approximativt oberoende normalfördelade med samma medelvärde men med olika varianser (ekvation 3.13 och 3.14).

$$\sigma_1^2 = \frac{1}{(n_1-3)} \tag{3.13}$$

$$\sigma_2^2 = \frac{1}{(n_2-3)} \tag{3.14}$$

Signifikanstest kan då göras med teststorheten $U$, där

$$U = \frac{Z_1 - Z_2}{\sqrt{\left(\frac{1}{n_1-3} + \frac{1}{n_2-3}\right)}}. \tag{3.15}$$

Om $\mid U \mid > 1.96$ (95 procents säkerhetsnivå) kan $H_0$ förkastas.

### 3.3.2 Korrelation för kategoriska variabler

När variablerna är kategoriska, kan data sorteras in i en korstabell. För varje par av variabler, finns det $n$ enheter kategoriserade i tabellen. Exempel på struktur finns i tabell 3.3.

Tabell 3.3: Exempel på korstabell för kategoriska variabler

| | | Grupp 2 | | |
|---|---|---|---|---|
| | | 1 | 0 | Totalt |
| Grupp 1 | 1 | a | b | a+b |
| | 0 | c | d | c+d |
| | Totalt | a+c | b+d | n=a+b+c+d |

Produktkorrelationskoefficienten beräknas då enligt [9] som

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{3.16}$$

### 3.3.3 Klusteranalyser

Vilket diskuterades i avsnitt 2.5.2.2 används i denna undersökning euklidiska avstånd vid beräkning av likheter (närhet) mellan objekt. Avståndet mellan punkt $i$ och punkt $j$ definieras enligt [10] som

$$d_{i,j} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}. \tag{3.17}$$

För matematisk definition av de klustermetoder som presenterades i 2.5.2.2 hänvisas till inbyggda hjälpfiler i R där den intresserade kan finna mer djupgående information om hur klusteranalyserna är strukturerade rent tekniskt[1].

## 3.4 Klassificeringar med regressionsanalys

I 3.2.1 presenterades logistisk regression och såsom också diskuterades kan denna modell används vid skapandet av klassificeringsmodeller. Om logaritmen av oddskvoten $\geq 0$ klassificeras den nya kunden till grupp 1 (se ekvation 3.19) och i annat fall till grupp 2. Detta kan utryckas som

$$\frac{\widehat{p}\,(z)}{1 - \hat{p}\,(z)} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 z_1 + \ldots + \hat{\beta}_r z_r\right) > 1. \tag{3.18}$$

Om vi använder den naturliga logaritmen på båda sidor erhålls

$$\ln \frac{\hat{p}(z)}{1 - \hat{p}(z)} = \hat{\beta}_0 + \hat{\beta}_1 z_1 + \ldots + \hat{\beta}_r z_r > 0. \tag{3.19}$$

Ett litet exempel kan vara på sin plats. Antag att vi är intresserade av att beräkna oddskvoten då sannolikheten $p$ befinner sig strax runt 50 procent:

- $p_1 = 0.55 \Rightarrow \frac{\widehat{p}(z)}{1-\hat{p}(z)} = \frac{0.55}{1-0.55} = 1.22$

[1] exempelvis `help(dist)`, `help(hclust)` eller `help(cutree)`

- $p_2 = 0.50 \Rightarrow \frac{\widehat{p}(z)}{1-\widehat{p}(z)} = \frac{0.50}{1-0.50} = 1.00$
- $p_3 = 0.45 \Rightarrow \frac{\widehat{p}(z)}{1-\widehat{p}(z)} = \frac{0.45}{1-0.45} = 0.82$

Alltså kan konstateras hur oddskvoten ger värden $\geq 1$ då $p \geq 0.5$. Logistisk regression kan alltså användas för att klassificera observationer till någon av de två grupperna. För varje specifik observationer går det också att erhålla en specifik procentsats för hur troligt det är att en observation hamnar i någon av de två grupperna (baserat på vilka värden variablerna antar).

### 3.4.1 Hur korrekt är klassificeringsmodellen?

Genom att antingen använda alla observationer som använts för att skapa klassificeringsmodellen, eller genom att använda en sparad testmängd (eng. *testingset*) kan modellens felfrekvenser testas. I tabell 3.4 återfinns en struktur för hur observationer kan delas in i fyra grupper (enligt predikterad tillhörighet och egentlig tillhörighet). Denna tabell kan sedan användas för att beräkna hur korrekt den aktuella modellen klassificerar (så kallat APER).

Tabell 3.4: Korstabell för beräkning av APER

| | | Predikterad grupp | |
|---|---|---|---|
| | | Grupp 1 | Grupp 2 |
| Egentlig grupp | Grupp 1 | a | b |
| | Grupp 2 | c | d |

Tabellen 3.4 kan användas för att beräkna APER (Apparent Error Rate) enligt

$$APER = \frac{b+c}{a+b+c+d} \times 100. \tag{3.20}$$

Denna procentsats anger feltolkningen i modellen och vill givetvis minimeras. I det extrema fallet då alla observationer klassas rätt är APER=0 procent.

# Kapitel 4

# Analys av data

## 4.1 Introduktion till data

Utöver de tre mer ingående analyserna som genomförs senare i detta kapitel (avsnitt 4.3-4.5), presenteras i detta introducerande avsnitt lite beskrivande statistik för den data[1] som ligger till grund för senare analyser och modeller.

Tabell 4.1: Beskrivande statistik för kontinuerliga variabler (Grupp 1)

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| Min | 1 | 0 | 0 | 19 | 46 | 34 304 |
| Kvartil 1 | 7 | 12 | 249 | 40 | 69 | 190 846 |
| Median | 14 | 17 | 249 | 48 | 75 | 293 141 |
| Medel | 14 | 15 | 377 | 48 | 74 | 313 385 |
| Kvartil 3 | 21 | 20 | 477 | 56 | 80 | 395 679 |
| Max | 31 | 23 | 995 | 80 | 100 | 3 560 816 |

Tabell 4.2: Beskrivande statistik för kontinuerliga variabler (Grupp 2)

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|---|
| Min | 1 | 0 | 179 | 19 | 0 | 48 891 |
| Kvartil 1 | 6 | 11 | 249 | 29 | 57 | 131 130 |
| Median | 11 | 15 | 249 | 42 | 69 | 251 812 |
| Medel | 14 | 14 | 431 | 40 | 66 | 251 339 |
| Kvartil 3 | 22 | 19 | 699 | 49 | 75 | 331 596 |
| Max | 31 | 23 | 995 | 84 | 90 | 713 054 |

I matriserna 4.1 och 4.2 presenteras de skattade korrelationskoefficienterna

[1] $X_1$ (Datum), $X_2$ (Tidpunkt), $X_3$ (Belopp), $X_4$ (Ålder), $X_5$ (Poäng), $X_6$ (Inkomst), $X_7$ (NyKund) och $X_8$ (Kön)

för respektive grupp.

$$\rho_1 = \begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ X_1 & 1 & & & & & & & \\ X_2 & -0.04 & 1 & & & & & & \\ X_3 & 0.06 & 0.01 & 1 & & & & & \\ X_4 & 0.06 & 0.03 & 0.01 & 1 & & & & \\ X_5 & 0.03 & 0.07 & 0.04 & 0.06 & 1 & & & \\ X_6 & 0.07 & 0.04 & 0.07 & 0.03 & 0.33 & 1 & & \\ X_7 & 0.07 & -0.02 & 0.07 & -0.06 & -0.03 & 0.09 & 1 & \\ X_8 & 0.07 & 0.05 & 0.13 & -0.04 & 0.10 & 0.25 & 0.13 & 1 \end{bmatrix} \tag{4.1}$$

$$\rho_2 = \begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ X_1 & 1 & & & & & & & \\ X_2 & 0.10 & 1 & & & & & & \\ X_3 & 0.03 & 0.08 & 1 & & & & & \\ X_4 & 0.20 & 0.12 & 0.16 & 1 & & & & \\ X_5 & 0.19 & 0.14 & -0.06 & 0.45 & 1 & & & \\ X_6 & 0.08 & 0.06 & 0.06 & 0.36 & 0.56 & 1 & & \\ X_7 & 0.00 & -0.04 & 0.13 & -0.04 & -0.12 & 0.01 & 1 & \\ X_8 & 0.03 & 0.07 & 0.15 & 0.00 & 0.13 & 0.24 & 0.05 & 1 \end{bmatrix} \tag{4.2}$$

I tabell 4.3-4.4 presenteras förekomster för de kategoriska variablerna. I tabellerna presenteras också sannolikheten för utfallet av den första kolumnens för respektive variabel.

Tabell 4.3: Korstabell för kategoriska variabeln NyKund

| | Sant | Falskt | Sannolikhet |
|---|---|---|---|
| Grupp 1 | 583 | 433 | 0.57 |
| Grupp 2 | 89 | 51 | 0.64 |

Tabell 4.4: Korstabell för kategoriska variabeln Kön

| | Man | Kvinna | Sannolikhet |
|---|---|---|---|
| Grupp 1 | 492 | 524 | 0.48 |
| Grupp 2 | 76 | 64 | 0.54 |

## 4.2 Något om algoritmer och simuleringar

I avsnitt 4.4-4.5 presenteras algoritmer för hur modeller konstruerats. Syftet med dessa modeller är att skapa en allmängiltig metodik för hur företagets kreditrisker kan analyseras. Med en allmängiltig modell blir det enklare att

genomföra analyser på andra datamängder än den som legat till grund för detta examensarbete.

Simuleringsstudier, eller Monte Carlo-metoder, används för att skaffa approximativ information om storheter (ofta sannolikheter eller väntevärden) som man inte klarar av att beräkna analytiskt [2]. Simulering innebär att, i största möjliga mån återskapa en verklighet. I detta fall är syftet med simuleringen att minska risken för att ett resultat är felaktigt till följd av slump. Genom att simulera en modell ett stort antal gånger (exempelvis 100 gånger) och sedan analysera dessa 100 resultat istället för att bara analysera 1 resultat minskar risken för felaktiga slutsatser. I [2] presenteras en allmän metodik för Monte-Carlo simuleringar:

1. Ställ upp en lämplig slumpmodell
2. Generera slumptal från de ingående fördelningarna
3. Beräkna värdet av den intressanta storheten
4. Upprepa steg 2 och 3 ett lämpligt (ofta stort) antal gånger
5. Skatta storheten med medelvärdet av de observerade värdena

Metodiken ovan fungerar som grund för de algoritmer som presenteras i 4.4-4.5.

I avsnitt 4.4.2 presenteras en algoritm för att genomföra klusteranalys av hela populationen. I algoritmen står hur 50 observationer slumpmässigt plockas ut från respektive grupp och att dessa 100 observationer sedan används för analyser. En risk med detta urval är att dessa 100 observationer inte är representativt för hela populationen (eftersom observationerna har valts slumpmässigt) och därför simuleras detta 1000 gånger för att på så sätt minimera risken för felaktiga slutsatser.

I flera av algoritmerna står också hur avståndsmatriser skapas och hur dessa sedan anpassas med lämplig klustermetodik. För mer ingående förklaringar till varför detta görs, se avsnitt 2.5.2.2.

## 4.3 Signifikanstester och hypotesprövningar

I tabell 4.5 presenteras resultaten av de parvisa signifikanstesterna som gjorts i R med hjälp av Mann-Whitneys test. Resultaten visar att för fem av totalt åtta variabler finns en signifikant skillnad i fördelning mellan de båda populationerna. Detta resultat kan jämföras med resultaten av vilka variabler som anses signifikanta i de regressionsmodeller som diskuteras i avsnitt 4.5.1. Rimligtvis borde de variablerna där skillnad föreligger sammanfalla med de variablerna som är signifikanta för regressionsmodellerna.

## 4.4 Korrelations- och klusteranalyser

Vid analys av korrelationsmatriserna 4.1 och 4.2 inses hur det finns tre variabelpar som ser ut att skilja sig mellan de båda populationerna. För grupp 2 är

Tabell 4.5: Beräkning av signifikans hos variabler med Mann-Whitneys test

| | $p$-värde | Återgärd | Resultat |
|---|---|---|---|
| Datum | 0.18 | Förkasta inte $H_0$ | Ingen skillnad föreligger |
| Tidpunkt | 0.0031 | Förkasta $H_0$ | Skillnad föreligger |
| Belopp | 0.016 | Förkasta $H_0$ | Skillnad föreligger |
| Ålder | 1.1e-11 | Förkasta $H_0$ | Skillnad föreligger |
| Poäng | 8.6e-13 | Förkasta $H_0$ | Skillnad föreligger |
| Inkomst | 4.9e-05 | Förkasta $H_0$ | Skillnad föreligger |
| NyKund | 0.19 | Förkasta inte $H_0$ | Ingen skillnad föreligger |
| Kön | 0.23 | Förkasta inte $H_0$ | Ingen skillnad föreligger |

korrelationerna för $X_4$ (Ålder), $X_5$ (Poäng), $X_6$ (Inkomst) betydligt högre än för grupp 1. Genom att använda Fishers Z- transformation (se avsnitt 3.3.1.1) för att testa signifikanta skillnader hos $\rho$ fås resultat enligt tabell 4.6 med $H_0\,(\rho_1 = \rho_2)$.

Tabell 4.6: Beräkning av signifikanta skillnader mellan korrelationskoefficienter med hjälp av Fishers Z-transform

| | $Z_1$ | $Z_2$ | $\mid U \mid$ | Återgärd | Resultat |
|---|---|---|---|---|---|
| $X_4 - X_5$ | 0.0600 | 0.4842 | 4.9458 | Förkasta $H_0$ | Skillnader föreligger |
| $X_5 - X_6$ | 0.3424 | 0.06321 | 3.3776 | Förkasta $H_0$ | Skillnader föreligger |
| $X_4 - X_6$ | 0.0300 | 0.3765 | 4.0398 | Förkasta $H_0$ | Skillnader föreligger |

Klusteranalyserna är uppdelade i två delar. I avsnitt 4.4.1 presenteras initialt en algoritm för klusteranalyser av grupp 2 och i underavsnitten 4.4.1.1-4.4.1.4 analyseras de kluster som träder fram mer ingående. I avsnitt 4.4.2 presenteras liknande algoritm för klusteranalys av hela populationen (observationer från båda grupperna).

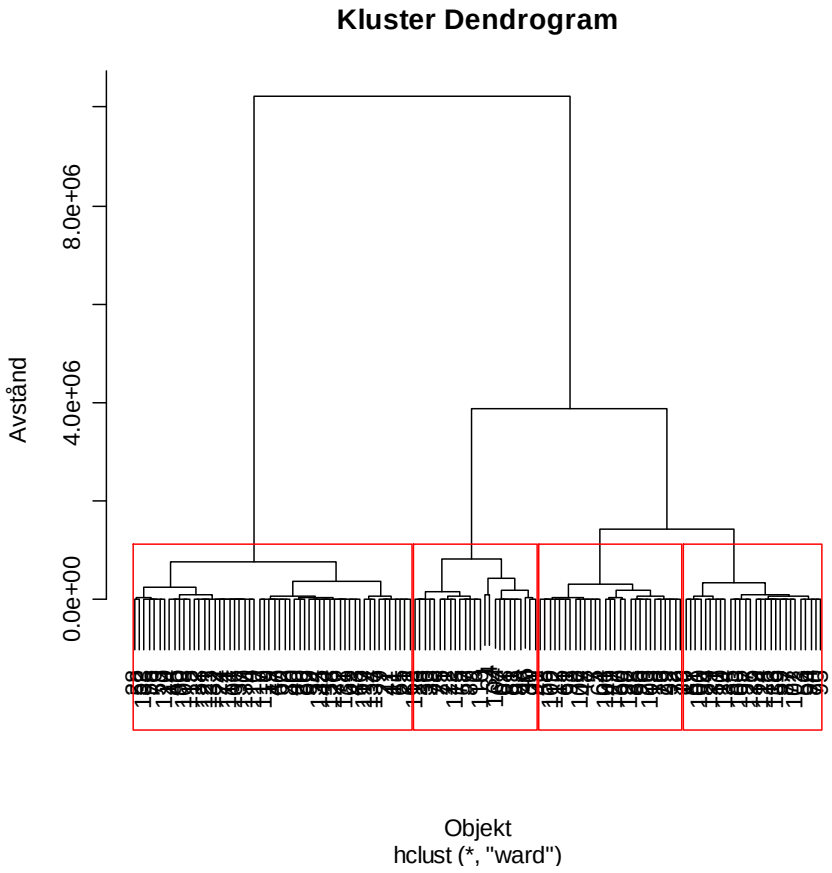
Genom att identifiera korrelationer inom respektive kluster förbättras insikterna i vad det är som utgör respektive kluster vilket i sig ger möjligheter (efter att detta testas på större populationer) till utarbetning av *reject-kriterier*.

### 4.4.1 Algoritm för klusteranalys av grupp 2

- Utgå från initial datamängd (med samtliga observationer) och sortera ur detta ut de observationer som skall ingå i klusteranalys (i detta fall valdes 138 observationerna från grupp 2). Dessa observationer samlas i en ny matris.

- För denna nya matris mäts likheten mellan observationerna med hjälp av godtycklig avståndsmetodik (exempelvis euklidiskt avstånd, se avsnitt 3.3.3). Dessa avstånd samlas i en ny matris (den så kallade avståndsmatrisen) vilken ligger till grund för skapandet av kluster.

- Anpassa (eng. *fit*) avståndsmatrisen med hjälp av godtycklig klustermetod (exempelvis Wards klustermetod, se avsnitt 3.3.3).
- Plotta anpassningen i ett dendrogram för att på så sätt möjliggöra identifiering av troliga kluster. Se exempelvis figur 4.1.
- Skapa ett antal nya tomma matriser (en för varje identifierat kluster, i detta fall fyra stycken) och sortera in observationerna för respektive kluster i dessa nya matriser. Nu finns för varje kluster en matris bestående av observationer (och dess tillhörande variabelvärden) för just detta kluster.
- Beräkna korrelationskoefficienterna för dessa matriser för att på så sätt identifiera vilka kombinationer av variabler som primärt utgör respektive klustret (se korrelationsmatriserna som inleder avsnitt 4.4.1.1-4.4.1.4).
- Notera de tre högsta korrelationerna för respektive kluster och plotta spridningsdiagram för dessa för att underlätta analys för respektive kluster.




Figur 4.1: Kluster dendrogram över observationer från grupp 2

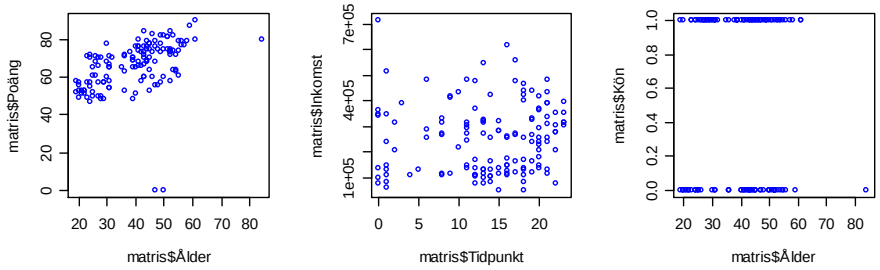
#### 4.4.1.1 Analys av korrelation för kluster 1

$$
\begin{bmatrix}
 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\
X_1 & 1 & & & & & & & \\
X_2 & -0.091 & 1 & & & & & & \\
X_3 & 0.126 & -0.288 & 1 & & & & & \\
X_4 & 0.322 & -0.072 & 0.023 & 1 & & & & \\
X_5 & 0.162 & -0.097 & -0.163 & 0.651 & 1 & & & \\
X_6 & 0.079 & -0.412 & 0.171 & 0.227 & 0.251 & 1 & & \\
X_7 & 0.266 & -0.120 & 0.159 & 0.099 & -0.075 & 0.069 & 1 & \\
X_8 & -0.269 & 0.158 & 0.210 & -0.401 & -0.351 & 0.033 & 0.081 & 1
\end{bmatrix}
$$

De tre största korrelationerna i detta kluster visar sig vara:

- $X_4 - X_5$: Ålder och Poäng ($\rho = 0.65$)

- $X_2 - X_6$: Tidpunkt och Inkomst ($\rho = -0.41$)
- $X_4 - X_8$: Ålder och Kön ($\rho = -0.40$)



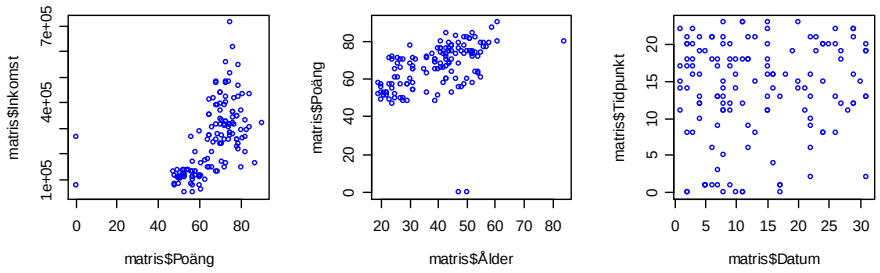

Figur 4.2: Spridningsdiagram för de tre högsta korrelationerna för kluster 1

#### 4.4.1.2 Analys av korrelation för kluster 2

$$\begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ X_1 & 1 & & & & & & & \\ X_2 & 0.278 & 1 & & & & & & \\ X_3 & 0.013 & 0.192 & 1 & & & & & \\ X_4 & 0.171 & 0.256 & 0.090 & 1 & & & & \\ X_5 & 0.127 & 0.165 & 0.112 & 0.376 & 1 & & & \\ X_6 & 0.107 & 0.216 & 0.064 & 0.137 & 0.482 & 1 & & \\ X_7 & -0.178 & -0.141 & 0.124 & 0.040 & -0.129 & -0.211 & 1 & \\ X_8 & 0.000 & -0.027 & 0.169 & 0.019 & -0.091 & 0.009 & 0.043 & 1 \end{bmatrix}$$

De tre största korrelationerna i detta kluster visar sig vara:

- $X_5 - X_6$: Poäng och Inkomst ($\rho = 0.48$)
- $X_4 - X_5$: Ålder och Poäng ($\rho = 0.38$)
- $X_1 - X_2$: Datum och Tidpunkt ($\rho = 0.28$)



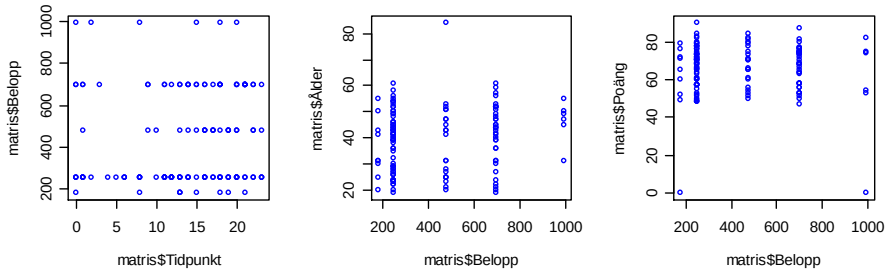

Figur 4.3: Spridningsdiagram för de tre högsta korrelationerna för kluster 2

#### 4.4.1.3 Analys av korrelation för kluster 3

$$\begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ X_1 & 1 & & & & & & & \\ X_2 & 0.093 & 1 & & & & & & \\ X_3 & 0.453 & -0.067 & 1 & & & & & \\ X_4 & 0.405 & -0.160 & 0.284 & 1 & & & & \\ X_5 & 0.149 & -0.253 & 0.031 & 0.678 & 1 & & & \\ X_6 & -0.007 & -0.202 & -0.054 & -0.061 & -0.156 & 1 & & \\ X_7 & 0.193 & 0.163 & 0.197 & -0.132 & -0.249 & -0.118 & 1 & \\ X_8 & -0.154 & 0.060 & 0.035 & -0.182 & -0.064 & 0.293 & -0.173 & 1 \end{bmatrix}$$

De tre största korrelationerna i detta kluster visar sig vara:

- $X_4 - X_5$: Ålder och Poäng ($\rho = 0.68$)
- $X_1 - X_3$: Datum och Belopp ($\rho = 0.45$)
- $X_1 - X_4$: Datum och Ålder ($\rho = 0.41$)



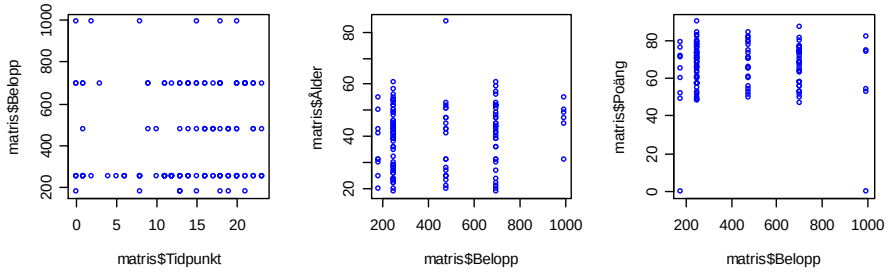

Figur 4.4: Spridningsdiagram för de tre högsta korrelationerna för kluster 3

#### 4.4.1.4 Analys av korrelation för kluster 4

$$\begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ X_1 & 1 & & & & & & & \\ X_2 & -0.132 & 1 & & & & & & \\ X_3 & -0.240 & 0.378 & 1 & & & & & \\ X_4 & -0.160 & 0.089 & 0.386 & 1 & & & & \\ X_5 & 0.084 & 0.245 & -0.345 & 0.053 & 1 & & & \\ X_6 & -0.215 & -0.011 & -0.117 & 0.091 & -0.125 & 1 & & \\ X_7 & -0.038 & 0.032 & 0.031 & -0.326 & -0.200 & -0.025 & 1 & \\ X_8 & 0.248 & 0.175 & 0.228 & -0.038 & 0.222 & -0.049 & 0.053 & 1 \end{bmatrix}$$

De tre största korrelationerna i detta kluster visar sig vara:

- $X_2 - X_3$: Tidpunkt och Belopp ($\rho = 0.38$)
- $X_3 - X_4$: Belopp och Ålder ($\rho = 0.39$)
- $X_3 - X_5$: Belopp och Poäng ($\rho = -0.35$)

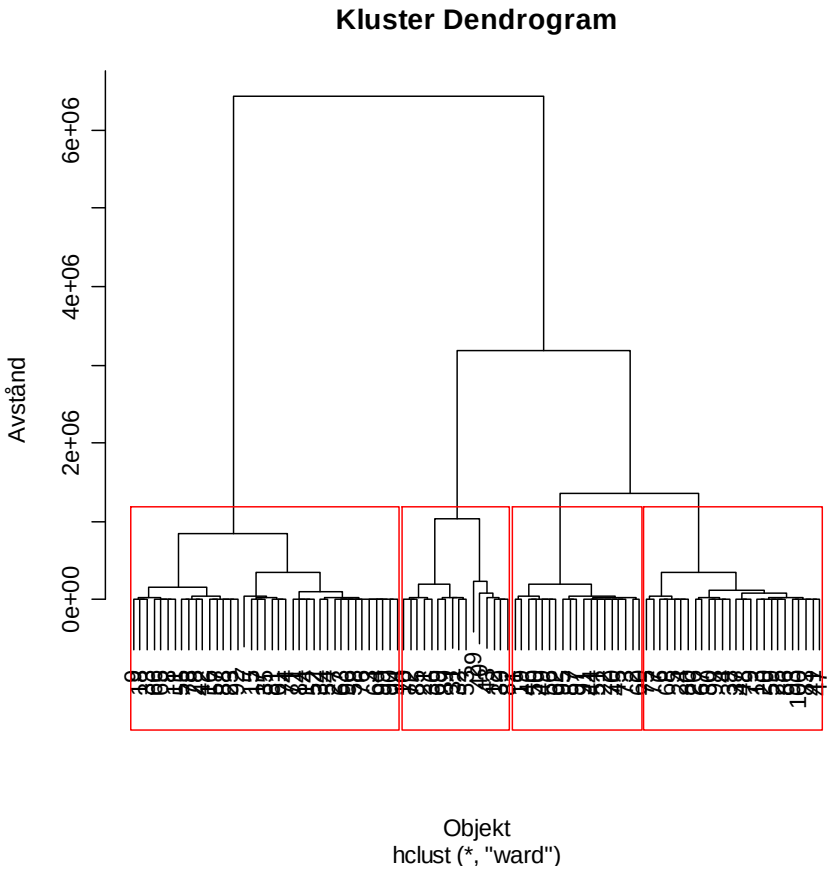



Figur 4.5: Spridningsdiagram för de tre högsta korrelationerna för kluster 4

### 4.4.2 Algoritm för klusteranalys av hela populationen

- Från initial datamängd plockas slumpmässigt 50 observationer ut från respektive grupp. Dessa 100 observationer (med dess variabelvärden) samlas i en ny matris.
- För denna nya matris mäts likheten mellan observationerna med hjälp av godtycklig avståndsmetodik (exempelvis euklidiskt avstånd, se avsnitt 2.5.2.2). Dessa avstånd samlas i en ny matris (den så kallade avståndsmatrisen) vilken ligger till grund för skapandet av kluster.
- Anpassa (eng. *fit*) avståndsmatrisen med hjälp av godtycklig klustermetod (exempelvis Wards klustermetod, se avsnitt 2.5.2.2).

- Plotta anpassningen i ett dendrogram för att på så sätt kunna identifiera rimliga kluster (se exempelvis figur 4.6).
- Skapa ett antal nya tomma matriser (en för varje identifierat kluster, i detta fall fyra stycken) och sortera in observationerna för respektive kluster i dessa nya matriser. Nu finns för varje kluster en matris bestående av observationer (och dess variabler) för respektive kluster.
- Notera hur många observationer från respektive grupp (grupp 1 eller grupp 2) som finns i respektive kluster.
- Simulera denna algoritm ett stort antal gånger (exempelvis 1000 gånger). För varje simulering, notera hur många observationer från respektive grupp som återfinns i respektive kluster.
- Addera samtliga observationer (när alla simuleringar är gjorda) från respektive grupp för varje kluster för alla simuleringar). Detta återfinns i tabell 4.7
- Beräkna kvoten för respektive kluster för att identifiera homogeniteten (se tabell 4.7).




Figur 4.6: Kluster dendrogram för en av de många simuleringarna med samtliga observationer

Tabell 4.7: Fördelning av observationer mellan de fyra grupperna för respektive kluster (hela populationen)

| | Kluster 1 | Kluster 2 | Kluster 3 | Kluster 4 |
|---|---|---|---|---|
| Grupp 1 | 14906 | 14266 | 12320 | 8508 |
| Grupp 2 | 14267 | 14057 | 12644 | 9032 |
| Kvot | 0.51/0.49 | 0.5/0.5 | 0.49/0.51 | 0.49/0.51 |

Det verkar alltså inte finnas någon tydlig skillnad mellan de båda grupperna vid ett stort antal simuleringar.

## 4.5 Klassificeringar med regressionsanalys

### 4.5.1 Algoritm för beräkning av signifikans hos variabler

- Från initial datamängd plockas slumpmässigt 100 observationer ut från respektive grupp. Dessa 200 observationer (med dess variabelvärden) samlas i en ny matris.
- Anpassa en logistisk regressionsmodell för dessa 8 variabler för den nya matrsien. I tabell 4.8 finns ett exempel på resultat från en sådan passning för en av simuleringarna.
- Notera vilka variabler som är signifikanta på 95 procents säkerhetsnivå. I exemplet (se tabell 4.8) framträder $X_3$ (Belopp), $X_4$ (Ålder) och $X_5$ (Poäng) som signifikanta på 95 procents säkerhetsnivå.
- Simulera detta ett stort antal gånger (exempelvis 100 gånger) och notera vilka variabler som är signifikanta för varje anpassning.
- Summera frekvenserna av respektive variabel (för alla simuleringar) enligt struktur i tabell 4.9

Tabell 4.8: Exempel på resultat från anpassning med logistisk regression

| | Coefficients | Estimated Std. Error | Z-Value | Pr (>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -5.539e+00 | 1.489e+00 | -3.719 | 0.00020 | *** |
| Datum | -5.971e-03 | 1.820e-02 | -0.328 | 0.74290 | |
| Tidpunkt | 3.661e-02 | 2.676e-02 | 1.368 | 0.17133 | |
| Belopp | -2.024e-03 | 8.586e-04 | -2.357 | 0.01840 | * |
| Ålder | 5.243e-02 | 1.714e-02 | 3.060 | 0.00222 | ** |
| Poäng | 4.353e-02 | 2.347e-02 | 1.855 | 0.06359 | . |
| Inkomst | 1.891e-06 | 1.320e-06 | 1.433 | 0.15191 | |
| NyKund | -8.341e-02 | 3.330e-01 | -0.251 | 0.80219 | |
| Kön | -1.821e-01 | 3.485e-01 | -0.523 | 0.60117 | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tabell 4.9: Förekomster av signifikanta variabler vid simulering 100 gånger

| Variabel | Frekvens |
|---|---|
| Ålder | 78 |
| Belopp | 69 |
| Poäng | 39 |
| Tidpunkt | 24 |
| Inkomst | 18 |
| Kön | 14 |
| Datum | 6 |
| NyKund | 4 |

### 4.5.2 Beräkning av signifikans hos variabler (hela populationen)

I 4.5.1 har ett slumpmässigt urval legat till grund för signifikanstester och konstruering av klassifcerare. Som ett komplement till detta anpassas i detta avsnitt en logistisk regressionsmodell där samtliga observationer används. Det kan vara intressant att prova hur utfallet skulle bli och jämföra det med resultaten från modelleringarna med 100 slumpmässiga observationer från respektive grupp. I tabell 4.10 presenteras en sammanställning över resultaten av modellering med samtliga observationer (alltså där kraftig skevhet förekommer, cirka 12 procent av observationerna kommer från grupp 2).

Tabell 4.10: Resultat från anpassning med logistisk regression (samtliga observationer)

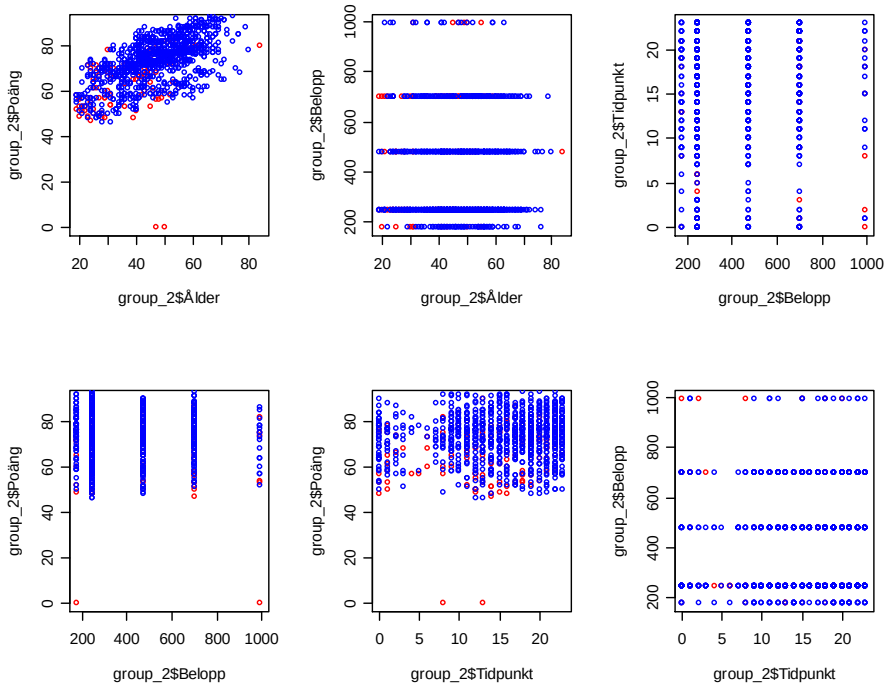
| | Coefficients | Estimated Std. Error | Z-Value | Pr (>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -2.131e+00 | 6.913e-01 | -3.083 | 0.00205 | ** |
| Datum | 6.412e-03 | 1.108e-02 | 0.579 | 0.56264 | |
| Tidpunkt | 3.222e-02 | 1.413e-02 | 2.279 | 0.02264 | * |
| Belopp | -1.298e-03 | 4.531e-04 | -2.865 | 0.00417 | ** |
| Ålder | 3.082e-02 | 9.582e-03 | 3.216 | 0.00130 | ** |
| Poäng | 3.723e-02 | 1.228e-02 | 3.031 | 0.00244 | ** |
| Inkomst | 1.168e-06 | 7.522e-07 | 1.553 | 0.12039 | |
| NyKund | -7.595e-02 | 1.994e-01 | -0.381 | 0.70325 | |
| Kön | -3.322e-01 | 2.028e-01 | -1.638 | 0.10136 | |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

En logistisk regressionsmodell (baserat på utfallet från skattningen i tabell 4.10) ansätts då som

$$Y = -2.13 + 0.0322 \times Tidpunkt - 0.00130 \times Belopp + 0.0308 \times Ålder + 0.0372 \times Poäng. \quad (4.3)$$

Spridningsdiagram för de signifikanta variablerna (utifrån resultat presenterat i tabell4.10) presenteras i figur 4.7.




Figur 4.7: Spridningsdiagram för de mest signifikanta variablerna (hela populationen)

### 4.5.3 Algoritm för beräkning av träffsäkerhet hos modell

- Från initial datamängd plockas slumpmässigt 100 observationer ut från respektive grupp. Dessa 200 observationer (med dess variabelvärden) samlas i en ny matris. Denna matris fungerar som träningsmängd (eng. *training-set*) för klassificeraren (se matris 4.4 för exempel på struktur, där $X_9$ indikerar grupptillhörighet).
- Anpassa en logistisk regressionsmodell för dessa 8 variabler.
- Från initialt dataset plockas slumpmässigt 50 nya observationer ut från respektive grupp. Dessa observationer fungerar som testmängd (eng. *testingset*).

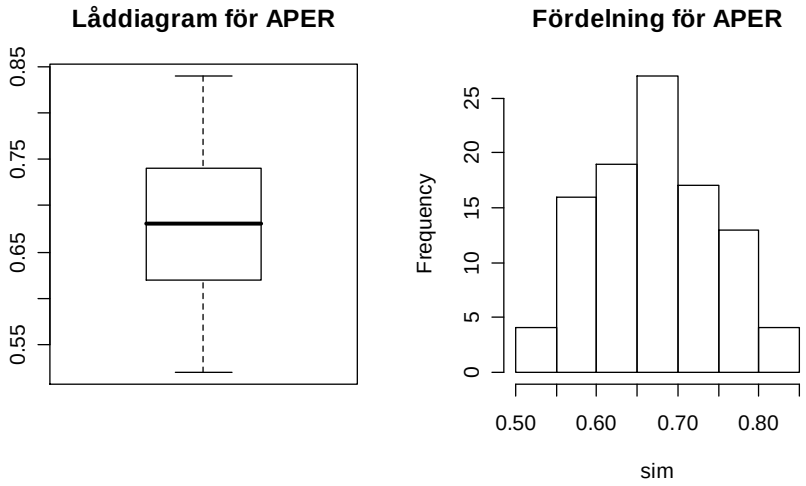
- Utifrån resultatet från den logistiska regressionsmodellen (vilket i statistikprogrammet presenteras som en tabell) beräknas prediktorvärdet för de 50 nya slumpmässigt valda observationerna. Beroende av vilken grupp vi vill testa väljs de 50 observationerna för denna grupp. De 50 prediktorvärden sparas i en vektor med 50 element.
- Utifrån teorin om "logistisk regression som klassificerare" (se avsnitt 3.4), vet vi att om ett prediktorvärde $\geq 0$ är observationen predikterad att tillhöra grupp 1. Studera på så sätt vektorn med de 50 predikterade elementen för respektive grupp för att avgöra om observationen blivit korrekt klassificerad.
- Skriv ner resultaten i en tabell (se struktur i tabell 4.12) och beräkna utifrån detta APER.
- Simulera detta ett stort antal gånger (exempelvis 100) för att konstatera hur bra modellen fungerar (se figur 4.8 för spridning och fördelning) .
- Beräkna APER för detta.

$$\begin{bmatrix} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 \\ 1 & 4 & 16 & 477 & 40 & 83 & 542590 & 0 & 0 & 1 \\ 2 & 6 & 7 & 249 & 57 & 75 & 156919 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 99 & 24 & 18 & 249 & 48 & 77 & 304711 & 0 & 1 & 1 \\ 100 & 9 & 12 & 249 & 39 & 54 & 42693 & 1 & 0 & 1 \\ 101 & 31 & 19 & 477 & 43 & 73 & 162428 & 0 & 1 & 0 \\ 102 & 12 & 18 & 249 & 20 & 57 & 50336 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 199 & 22 & 16 & 477 & 84 & 80 & 144719 & 1 & 0 & 0 \\ 200 & 8 & 20 & 249 & 54 & 64 & 180369 & 1 & 1 & 0 \end{bmatrix} \tag{4.4}$$

Tabell 4.12: Beräkning av APER för klassificerare

| | | Predikterad grupp | |
|---|---|---|---|
| | | Grupp 1 | Grupp 2 |
| Egentlig grupp | Grupp 1 | 32 | 17 |
| | Grupp 2 | 18 | 33 |

$$APER = 35\%$$




Figur 4.8: Fördelning för APER- beräkningar vid 100 simuleringar

Det kan konstateras hur den korrekta (och felaktiga) klassificeringen för de båda grupperna är ungefär lika stor, vilket med största sannolikhet har att göra med att gruppernas storlekar är lika stora i modellen (100 observationer från respektive grupp). Det konstateras i artiklar såsom exempelvis [6] att de båda grupperna måste vara lika stora vid konstrueringar av klassificerare. Att kunna klassificera med en felmarginal runt 35 procent får dock anses relativt dåligt.

# Kapitel 5

# Resultat och diskussion

## 5.1 Signifikanstester och hypotesprövningar

Resultat från både signifikanstester och regressionsanalyser visar på otvetydiga resultat. En variabel som inte skiljer sig märkvärt i fördelning mellan de båda grupperna bidrar inte heller med information i en regressionsmodell.

Vid signifikanstester (för test av skillnader i fördelning) framträder $X_2$ (Tidpunkt), $X_3$ (Belopp), $X_4$ (Ålder), $X_5$ (Poäng), $X_6$ (Inkomst) som signifikanta och vid regressionsanalyser (både med slumpmässiga stickprov samt med samtliga observationer) accepteras variablerna $X_2$ (Tidpunkt), $X_3$ (Belopp), $X_4$ (Ålder) och $X_5$ (Poäng).

Dessa resultat är inte särskilt förvånande då även [8] i sin studie konstaterat hur dessa variabler (samt några andra som inte ingått i denna studie) var signifikanta vid bedömning av kreditvärdighet. [8] fokuserade dock på en annan geografisk marknad och resultaten från deras studie i kombination med denna studie får anses styrka varandra beträffande vilka variabler som är relevanta att beakta vid bedömning av kreditvärdighet.

Vid både signifikanstester och regressionsanalyser förkastas de båda kategoriska variablerna (NyKund och Kön). I tabell 4.3-4.4 framgår det tydligt hur det inte verkar finnas någon skillnad mellan grupperna för dessa två variabler.

## 5.2 Korrelations- och klusteranalyser

Ett intressant samband som har identifierats är hur det finns tydliga samvariation mellan variablerna $X_4$ (Ålder), $X_5$ (Poäng) och $X_6$ (Inkomst) för grupp 2 medan dessa korrelationer är nästintill obefintliga för grupp 1. Att de är relaterade är naturligt då som tidigare diskuterats att $X_5$ (Poäng) till stor del bygger på de båda andra variablerna, men att detta samband inte existerar för grupp 1 intressant för vidare analys (se avsnitt 6.2 för vidare diskussion).

Efter att ha analyserat utfallen från klusteranalyserna kan konstateras hur det finns ett genomgående problem vilket kan kopplas till de initiala korrelation-

erna inom respektive grupp (se 4.1 och 4.2). Stark korrelation definieras som korrelationskoefficienter $\geq| 0.5 |$ och i de initiala korrelationsmatriserna uppfylls detta endast för grupp 2 där variablerna $X_5$ (Poäng) och $X_6$ (Inkomst) har en positiv korrelation ($\rho = 0.56$) med varandra. Med så förhållandevis svaga korrelationer, tillsammans med ett begränsat antal variabler och observationer, blir det svårt att identifiera några tydliga kluster.

I figur 4.1 återfinns ett dendrogram över klustertendenser för grupp 2. Fyra tänkbara kluster framträder i figuren och vid en närmare analys av inbördes korrelationer mellan observationer för respektive kluster (se avsnitt 4.4.1.1-4.4.1.4) inses hur även dessa korrelationer är för låga för att kunna säkerställa klustertendenser. Om korrelationsmatriserna för respektive kluster jämförs med korrelationsmatriserna för de ursprungliga grupperna kan konstateras hur de senare korrelationskoefficienterna är betydligt högre (med undantag för $X_5$, poäng).

De högsta korrelationerna i respektive kluster är kopplade till variabeln $X_5$ (Poäng) vilket också är den variabeln som idag används som huvudsakligt *reject-kriterium*. Då en del av syftet med denna studie är att utveckla nya *reject-kriterier* är det intressant att exkludera denna variabel ($X_5$) från klusteranalyserna och istället fokusera på de andra variablerna (se punkter nedan).

- Kluster 1 utgörs primärt av $X_2$ (Tidpunkt) och $X_6$ (Inkomst) där $\rho = -0.41$
- Kluster 2 utgörs primärt av $X_1$ (Datum) och $X_2$ (Tidpunkt) där $\rho = 0.28$
- Kluster 3 utgörs primärt av $X_1$ (Datum) och $X_3$ (Belopp) där $\rho = 0.45$ samt $X_1$ (Datum) och $X_4$ (Ålder) där $\rho = 0.41$
- Kluster 4 utgörs primärt av $X_1$(Datum) och $X_3$ (Belopp) där $\rho = 0.38$ samt $X_3$ (Belopp) och $X_4$ (Ålder) där $\rho = 0.39$

Om dessa korrelationskoefficienter jämförs med de initiala i 4.2 så kan konstateras att dessa koefficienter är avsevärt mycket högre vilket indikerar på att klusteranalyserna uppfyllt syftet med att föra samman liknande observationer utifrån dess egenskaper.

I figur 4.6 presenterades ett dendrogram med slumpmässiga observationer från båda grupper. Vid analys av tabell 4.7 inses dock hur det inte verkar finnas några klustertendenser vid ett stort antal simuleringar.

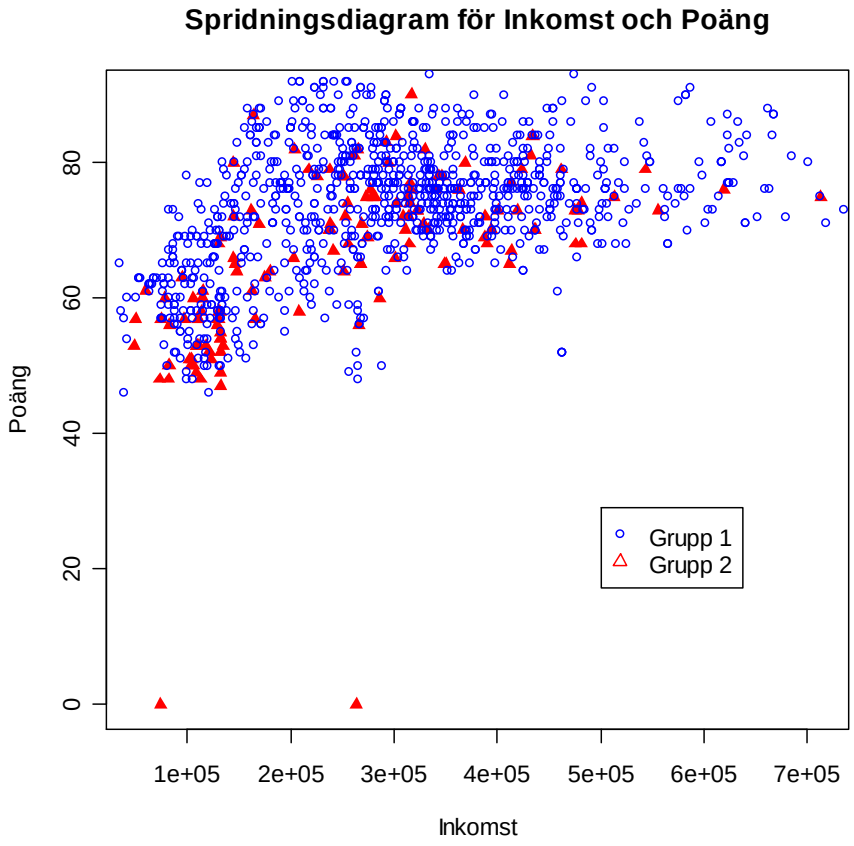
Detta stärker resultaten från klusteranalysen för grupp 2 där det konstaterades hur de ingående korrelationerna är för låga för att kunna genomföra relevanta klusteranalyser. Resultaten från klusteranalys bör därmed ses med viss försiktighet.

## 5.3 Klassificeringar med regressionsmodeller

I tabell 4.12 presenterades resultatet från klassificeringsmodell vid test på testmängden (eng. *testingset*). Den modellen som konstruerats klassificerar kunder-

na rätt i 65 procent av fallen. I 35 procent av fallen sker dock en felklassificering vilken är ungefär lika stor för båda grupperna.

I figur 5.1 återfinns ett spridningsdiagram för variablerna $X_5$ och $X_6$ (vilka är de variabler med högst korrelationer, se avsnitt 5.2) för respektive grupp plottade. I figuren syns tydligt hur observationerna från båda grupper går i varandra och hur det därmed blir svårt att dra någon klar avgränsare mellan grupperna (varken linjär eller icke-linjär). De båda grupperna är helt enkelt för lika varandra för att kunna konstruera en träffsäker klassificerare.




Figur 5.1: Spridningsdiagram för $X_5$ (Poäng) och $X_6$ (Inkomst).

# Kapitel 6

# Slutsatser

## 6.1 Signifikanstester och hypotesprövningar

- Resultat från regressionsanalyser och hypotesprövningar överensstämmer till stor del med de resultat som presenterades i [8]. Då [8] fokuserar på en annan geografisk marknad får de båda uppsatserna anses bekräfta varandras resultat. De variablerna som anses viktiga vid bedömning av trolig betalningsförmåga är alltså i stort desamma oavsett geografisk marknad.

## 6.2 Korrelations- och klusteranalyser

- Variablerna $X_4$ (Ålder), $X_5$ (Poäng) och $X_6$ (Inkomst) har svag (alternativt stark) korrelation för grupp 2 medan några sådana korrelation inte förekommer i grupp 1. Utifrån detta inses hur personer i grupp 1 tenderar ha många andra bra egenskaper (sett ur ett perspektiv för trolig betalningsförmåga) än just dessa variabler. För grupp 2 finns dock en tydlig homogenitet för dessa variabler och det är dessa variabler som är viktigast. En möjlig tolkning av detta kan kopplas till begreppen betalningsvilja och betalningsförmåga. En äldre person, med höga poäng och hög inkomst har statistiskt sett bra betalningsförmåga men betalningsviljan kan vara sämre. Personen kan helt enkelt vara lat, glömsk eller upptagen av annat än att betala skulder. Ett tänkbart *reject-kriterium* för detta skulle kunna vara att sätta en gräns för hur hög korrelationskoefficienten för dessa variabler tillåts vara för att en kund ska beviljas kredit.

- I kluster 3 konstateras hur en riskgrupp är äldre personer som ansöker om höga krediter i slutet av månaden. Detta bör undersökas på större datamängder för att undersöka om denna variabelkombination är återkommande. Om så är fallet skulle detta kunna implementeras som rimligt *reject-kriterium.*

- I exempelvis figur 5.1 finns två outliers där kunder med väldigt låga poäng ändå tillåtits krediter. Detta är kritiskt då dessa kunder enligt de regler som finns för kreditgivningen inte ska tillåtas några krediter överhuvudtaget. Detta är återkommande vid analys av olika dataset, och förekommer endast hos observationer från grupp 2.
- Klusteranalys är en bra analysmetod för att analysera multivariata statistiska samband vid bedömning av kreditvärdighet. Med hjälp av klusteranalyser är det enkelt att gruppera observationer och på så sätt utveckla nya tänkbara *reject-kriterier*. Nackdelen är att det behövs ett ganska stort antal observationer och variabler för att säkerställa kvalitén i dessa klusterbildningar. Då korrelationskoefficienterna i de fyra klustren är högre än de initiala korrelationsmatriserna finns goda belägg för att klusteranalysen fungerar bra för detta ändamål.
- Den viktigaste anledningen till att klusteranalysernas inte håller någon högre kvalitet kan kopplas till låga de korrelationerna i initialt dataset. En anledning till detta är det faktum att antalet variabler som fanns att tillgå var starkt begränsat. Hade fler variabler varit tillgängliga skulle intressantare klustertendenser (och *reject-kriterier*) kunna identifierats för respektive grupp.
- Variabeln $X_5$ (Poäng) är den variabeln som genomgående ingår i de flesta av de höga korrelationskoefficienterna som diskuterats. Denna variabel är vid samtliga analyser en av de viktigaste för bedömning av kreditvärdighet. Variabel fungerar idag också uteslutande som primärt *reject-kriterium*. Detta val känns högst relevant och befogat.

## 6.3 Klassificeringar med regressionsanalyser

- För att konstruera en så bra klassificerare som möjligt vore det bästa att som utgångspunkt genomföra korrelationsanalyser mellan samtliga variabler och sedan välja ut variabler med korrelationskoefficienter mellan exempelvis $\mid 0.3 \mid \leq \rho < \mid 0.5 \mid$. Att genomföra korrelationsanalyser med ett stort antal möjliga variabler var dock ej möjligt i denna studie till följd av svårigheter att få tillgång till allt datamaterial om varje observation.
- De båda grupperna (grupp 1 och grupp 2) är för lika varandra för att kunna genomföra bra klassificeringar. Det går att identifiera tendenser, men säkerheten i klassificeringarna är för dåliga för att kunna använda en sådan modell vid kreditgivning.
- Att kunna identifiera en kund som troligtvis inte kommer betala är viktigt men vad som är minst lika viktigt är att kunna identifiera en kund som kommer betala sina skulder. Att ge avslag till en kund som kommer betala är kritiskt och det är därför är det extremt viktigt att noggrant testa potentiella *reject-kriterier* åt båda hållen innan de implementeras fullt ut.

# Kapitel 7

# Lärdomar och förslag på vidare undersökning

Vilket diskuterats vid flera tillfällen i detta examensarbete är de initiala korrelationerna mellan variabler för respektive grupp väldigt låga. Det betyder att det är svårt att identifiera tydliga tendenser om inte finns ett stort antal observationer och variabler finns att tillgå. Sambanden mellan saker och ting är helt enkelt för svaga. Problemet med ett stort antal variabler är å andra sidan att kvaliteten i analyserna försämmras. Att använda observationer från ett helt år skulle ge många observationer men det skulle bli väldigt allmängiltigt och risken att missa exempelvis säsongsvariationer ökar.

Mitt förslag på vidare undersökning är att exempelvis diskretisera kontinueliga variabler i godtyckliga intervall och sedan analysera dessa intervall mer djupgående. Varibeln Ålder ($X_4$) kan exemplevis delas in i kunder som är äldre eller yngre än 20. Denna nya och utbyggda data-matris kan sedan används precis som tidigare för regressionsmodeller, klusteranalyser och klassificeringar. Skillnaden är bara att korrelationsmatriserna förmodligen kommer blir tydligare och senare analyser kommer därmed bli mer användbara och träffsäkra. Precis som i denna studie presenteras i [8] hur exempelvis variabeln tidpunkt har betydelse men det förs ingen djupare diskussion om vilka tidsperioder det är som är av störst betydelse. Detta borde kunna analyseras djupare med denna metodik.

För att exemplifiera detta kan matris 7.1 studeras. Här återfinns tre slumpmässiga observationer från grupp 2 med tillhörande värden på för de åtta variablerna[1]

$$\begin{array}{r|cccccccc} & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \\ Obs_1 & 18 & 19 & 249 & 19 & 80 & 190000 & 1 & 0 \\ Obs_2 & 13 & 02 & 699 & 35 & 85 & 230000 & 1 & 0 \\ Obs_3 & 06 & 07 & 477 & 42 & 71 & 310000 & 0 & 1 \end{array} \tag{7.1}$$

För att bygga ut matris 7.1, används godtyckliga indelningar och tabell 7.1 visar ett förslag (med detta förslag byggs de initiala 8 variablerna ut till 11).

[1] $X_1$ (Datum), $X_2$ (Tidpunkt), $X_3$ (Belopp), $X_4$ (Ålder), $X_5$ (Poäng), $X_6$ (Inkomst), $X_7$ (NyKund) och $X_8$ (Kön)

Tabell 7.1: Tabell över tänkbara delintervall för variabler

| Variabel | Krav |
|---|---|
| $X_1$ : | Ansökan om kredit gjord mellan 1-15 dagen i månaden |
| $X_{2.1}$ : | Ansökan om kredit gjord mellan klockan 00-06 |
| $X_{2.2}$ : | Ansökan om kredit gjord mellan klockan 07-23 |
| $X_3$ : | Belopp för ansökan om kredit $< 250$ |
| $X_4$ : | Ålder hos kund $< 25$ |
| $X_{5.1}$ : | $70 \leq$ Poäng för kund $< 80$ |
| $X_{5.2}$ : | $80 \leq$ Poäng för kund $< 90$ |
| $X_{5.3}$ : | $90 \leq$ Poäng för kund |
| $X_6$ : | Inkomst hos kund $< 200\,000$ |
| $X_7$ : | Kundens kön är Man |
| $X_8$ : | Kunden är Ny |

Om villkoren för respektive variabel i tabell 7.1 uppfylls tilldelas det elementet en 1:a i den nya matrisen (se matris 7.2), annars en 0:a. På så sätt erhålls en ny data-matris bestående av 1:or och 0:or enligt strukturen i matris 7.2.

$$\begin{array}{r|cccccccccc} & X_1 & X_{2.1} & X_{2.2} & X_3 & X_4 & X_{5.1} & X_{5.2} & X_{5.3} & X_6 & X_7 & X_8 \\ Obs_1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ Obs_2 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ Obs_3 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array} \tag{7.2}$$

Med denna utbyggda data-matris finns bättre förutsättningar för att identifiera särskilt kritiska korrelationer som sedan kan användas som *rejet-kriterier* För att kunna utveckla träffsäkra *reject-kriterier* krävs dock signifikans och det är svårt att identifiera 100 procentiga *reject-kriterier* med de förhållandevis få observationer som ingår i denna studie.

En viktig aspekt att ha i åtanke är också att denna studie genomförts för en specifik kundgrupp på ett geografiskt avgränsat område. Innan slutsatser från denna uppsats kan implementeras krävs att dessa testas på större populationer och på fler markander för att säkerställa kvalitén i slutsatserna.

Som figur 2.1 visar kan en kreditskuld befinna sig i olika stadier. Denna uppsats har bara analyserat två grupper av dessa många möjliga gruppindelningar. Det är inget som säger att just de två grupper som analyserats i denna uppsats är det bästa valet. För att kunna analysera vad en "dålig" respektive "bra" grupp är krävs dock en mer djupgående analys av lönsamhet hos kunder.

# Litteraturförteckning


[1] A. Agresti (2007), *An Introduction to Categorical Data Analysis*, John Wiley & Sons, New Jersey

[2] S.E. Alm, T. Britton (2008), *Stokastik - Sannolikhetsteori och statistikteori med tillämpningar*, Liber, Stockholm

[3] G. Blom, B. Holmquist (1998), *Statistikteori med tillämpningar* (Bok B, 3e upplagan), Studentlitteratur, Lund

[4] P. Crosbie, J. Bohn (2003), *Modelling default risk,* Moody's KMV Corporation

[5] C. Edling, P. Hedström (2003), *Kvantitativa metoder*, Studentlitteratur, Lund

[6] M. A. Gouvêa (2007), *Credit risk analysis applying logistic regression, Neural networks and nenetic algorithms models*, University of São Paulo, Brazil

[7] T. Hastie, R. Tibshirani, J. Friedman (2002), *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, Springer-Verlag, New York

[8] V. Jacobsson, S. Siemiatkowski (2007), *Consumption credit default predictions*, Handelshögskolan i Stockholm, Stockholm

[9] R. A. Johnson, D. W. Wichern (2007), *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey

[10] W.J. Kranzanowski (1988), *Principles of Multivariate Analysis*, Oxford University Press, Oxford

[11] N.L Leech, K.C Barett, G.A Morgan (2008), *SPSS for Intermediate Statistics - Use and Interpretation*, Taylor & Francis Group, New York

[12] Riksgäldens styrelse (2009), *Finans- och riskpolicy 2009*, Riksgälden, Stockholm

[13] S. Siegel, N.J. Castellan (1988), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Co., Singapore

# Bilaga A

# Programkod för R

## A.1 Signifikanstester och hypotesprövningar

```
##### Inläsning av data och indelning i Grupp1 & Grupp2 #####
   data <- read.table("data2.dat", header=TRUE)
   group_1 <- matrix(c(1:(1016*9)),ncol=9, byrow=TRUE)
   group_2 <- matrix(c(1:(140*9)),ncol=9, byrow=TRUE)


   colnames(group_1) <-c("Datum","Tidpunkt","Belopp","Ålder",
   "Poäng","Inkomst","NyKund","Kön","Grupp")
   colnames(group_2) <-c("Datum","Tidpunkt","Belopp","Ålder",
   "Poäng","Inkomst","NyKund","Kön","Grupp")


   for(j in 1:1016){
   for(k in 1:9){
   group_1[j,k] <- data[j,k]
   }}
   for(j in 1:140){
   for(k in 1:9){
   group_2[j,k] <- data[1016+j,k]
   }}


   group_1 <- as.data.frame(group_1)
   group_2 <- as.data.frame(group_2)


   cor(group_1)
   cor(group_2)


   ##### Plot med spridningsdiagram #####
```

```
par(mfrow=c(2,3))
plot(group_2$Ålder,group_2$Poäng, type="p", col="red")
points(group_1$Ålder,group_1$Poäng,type="p",col="blue")
plot(group_2$Ålder,group_2$Belopp, type="p", col="red")
points(group_1$Ålder,group_1$Belopp,type="p",col="blue")
plot(group_2$Belopp,group_2$Tidpunkt, type="p", col="red")
points(group_1$Belopp,group_1$Tidpunkt,type="p",col="blue")
plot(group_2$Belopp,group_2$Poäng, type="p", col="red")
points(group_1$Belopp,group_1$Poäng,type="p",col="blue")
plot(group_2$Tidpunkt,group_2$Poäng, type="p", col="red")
points(group_1$Tidpunkt,group_1$Poäng,type="p",col="blue")
plot(group_2$Tidpunkt,group_2$Belopp, type="p", col="red")
points(group_1$Tidpunkt,group_1$Belopp,type="p",col="blue")
close.screen(all = TRUE)


plot(group_2$Ålder,group_2$Poäng, col="red", pch=17,
xlab="Ålder", ylab="Poäng",
main="Spridningsdiagram Ålder och Poäng")
points(group_1$Ålder,group_1$Poäng, col="blue", pch=1)
plot(group_2$Inkomst,group_2$Poäng, col="red", pch=17,
xlab="Inkomst", ylab="Poäng",
main="Spridningsdiagram Inkomst och Poäng")
points(group_1$Inkomst,group_1$Poäng, col="blue", pch=1)


##### Hypotesprövningar #####
### Chi2-Test ###
data <- matrix(c(145,128,30,21),nrow=2)
chisq.test(data)


### Wilcox/Man-Whitney U ###
data <- read.table("data.dat", col.names = c("Svar","Typ"),
header=TRUE)
data$Typ <- factor(data$Typ, labels=c("OK","INKASSO"))
wilcox.test(Svar~Typ,data, paired = FALSE)


##### Plot för histogram #####
split.screen(c(1,4))
screen(1)
hist(group_1$Tidpunkt, main="Tidpunkt (Grupp 1)")
screen(2)
hist(group_1$Poäng, main="Poäng (Grupp 1)")
screen(3)
hist(group_2$Tidpunkt, main="Tidpunkt (Grupp 2)")
screen(4)
```

```
   hist(group_2$Poäng, main="Poäng (Grupp 2)")
```

## A.2 Korrelations - och klusteranalyser (grupp 2)

```
##### Inläsning av data #####
   data <- read.table("data.dat", header=TRUE)
   matris <- matrix(c(1:(140*9)),ncol=9, byrow=TRUE)
   colnames(matris) <- c("Datum","Tidpunkt","Belopp","Ålder",
   "Poäng","Inkomst","NyKund","Kön","Grupp")


   for(j in 1:140){
   for(k in 1:9){
   matris[j,k] <- data[j+1016,k]
   }}

   matris <- as.data.frame(matris)
   matris2 <- matrix(c(1:(140*8)),ncol=8, byrow=TRUE)
   colnames(matris2) <- c("Datum","Tidpunkt","Belopp","Ålder",
   "Poäng","Inkomst","NyKund","Kön")
   for(j in 1:8){
   matris2[col(matris2) == j] <- matris[col(matris) == j]
   }
   matris2 <- as.data.frame(matris2)


   ##### Förberedelser för klusteranalys #####
   d <- dist(matris2, method="euclidean")
   fit <- hclust(d, method="ward")
   plot(fit)
   groups <- cutree(fit, k=4)
   rect.hclust(fit, k=4, border="red")

   ##### Klusteranalys #####
   clust1 <- as.data.frame(matrix(NA,700,ncol=8))
   clust2 <- as.data.frame(matrix(NA,700,ncol=8))
   clust3 <- as.data.frame(matrix(NA,700,ncol=8))
   clust4 <- as.data.frame(matrix(NA,700,ncol=8))


   colnames(clust1) <- c("Datum","Tidpunkt","Belopp",
   "Ålder","Poäng","Inkomst","NyKund","Kön")
   colnames(clust2) <- c("Datum","Tidpunkt","Belopp",
   "Ålder","Poäng","Inkomst","NyKund","Kön")
   colnames(clust3) <- c("Datum","Tidpunkt","Belopp",
   "Ålder","Poäng","Inkomst","NyKund","Kön")
   colnames(clust4) <- c("Datum","Tidpunkt","Belopp",
   "Ålder","Poäng","Inkomst","NyKund","Kön")
```

```
   for(k in 1:140){
   if (groups[k] == 1){
   clust1[row(clust1) == k*5] <- matris2[row(matris2) == k]
   }
   if (groups[k] == 2){
   clust2[row(clust2) == k*5] <- matris2[row(matris2) == k]
   }
   if (groups[k] == 3){
   clust3[row(clust3) == k*5] <- matris2[row(matris2) == k]
   }
   if (groups[k] == 4){
   clust4[row(clust4) == k*5] <- matris2[row(matris2) == k]
   }}


   clust1 <- na.omit(clust1);cor(clust1)
   clust2 <- na.omit(clust2);cor(clust2)
   clust3 <- na.omit(clust3);cor(clust3)
   clust4 <- na.omit(clust4);cor(clust4)

   ##### Plottning av de högsta korrelationerna #####
   par(mfrow=c(1,3))
   plot(matris$Tidpunkt,matris$Belopp, type="p", col="blue")
   plot(matris$Belopp,matris$Ålder, type="p", col="blue")
   plot(matris$Belopp,matris$Poäng, type="p", col="blue")
   close.screen(all = TRUE)
```

## A.3 Korrelations- och klusteranalyser (båda grupperna)

```
##### Inläsning av data #####
   data <- read.table("data.dat", header=TRUE)
   no <- 1000
   sim <- c(1:no)
   count_total_1 <- 0
   count_total_2 <- 0
   for(m in 1:no){


   ix1 <- sample(c(1:1016),50,replace=FALSE)
   ix2 <- sample(c(1017:1156),50,replace=FALSE)


   matris <- matrix(c(1:(100*9)),ncol=9, byrow=TRUE)
   colnames(matris) <- c("Datum","Tidpunkt","Belopp","Ålder",
   "Poäng","Inkomst","NyKund","Kön","Grupp")
```

```
for(j in 1:50){
for(k in 1:9){
matris[j,k] <- data[ix1[j],k]
matris[j+50,k] <- data[ix2[j],k]
}}


matris <- as.data.frame(matris)
matris2 <- matrix(c(1:(100*8)),ncol=8, byrow=TRUE)
colnames(matris2) <- c("Datum","Tidpunkt","Belopp","Ålder",
"Poäng","Inkomst","NyKund","Kön")
for(j in 1:8){
matris2[col(matris2) == j] <- matris[col(matris) == j]
}
matris2 <- as.data.frame(matris2)


##### Klusteranalys #####
d <- dist(matris2, method="euclidean")
fit <- hclust(d, method="ward")
plot(fit)
groups <- cutree(fit, k=4)
rect.hclust(fit, k=4, border="red")


group_1 <- groups[1:50]
group_2 <- groups[51:100]

clust1_1<-0; clust2_1<-0; clust3_1<-0; clust4_1<-0
clust1_2<-0; clust2_2<-0; clust3_2<-0; clust4_2<-0


for(k in 1:50){
if (group_1[k] == 1) {clust1_1 <-clust1_1+1}
if (group_2[k] == 1) {clust1_2 <-clust1_2+1}
if (group_1[k] == 2) {clust2_1 <-clust2_1+1}
if (group_2[k] == 2) {clust2_2 <-clust2_2+1}
if (group_1[k] == 3) {clust3_1 <-clust3_1+1}
if (group_2[k] == 3) {clust3_2 <-clust3_2+1}
if (group_1[k] == 4) {clust4_1 <-clust4_1+1}
if (group_2[k] == 4) {clust4_2 <-clust4_2+1}
}


count_1 <- c(clust1_1,clust2_1,clust3_1,clust4_1)
count_2 <- c(clust1_2,clust2_2,clust3_2,clust4_2)
count_total_1 <- count_total_1 + count_1
count_total_2 <- count_total_2 + count_2
}
```

```
   count_total_1
   count_total_2
```

## A.4 Klassificeringar med regressionsanalys

```
##### Inläsning av data #####
   data <- read.table("data.dat", header=TRUE)
   no <- 100
   sim <- c(1:no)
   for(m in 1:no){


   ##### Traingset #####
   ix1 <- sample(c(1:1016),100,replace=FALSE)
   ix2 <- sample(c(1017:1156),100,replace=FALSE)
   ##### Testingset #####
   ix3 <- sample(c(1:1016),50,replace=FALSE)
   ix4 <- sample(c(1017:1156),50,replace=FALSE)


   matris <- matrix(c(1:1800),ncol=9, byrow=TRUE)
   colnames(matris) <- c("Datum","Tidpunkt","Belopp","Ålder",
   "Poäng","Inkomst","NyKund","Kön","Grupp")


   for(j in 1:100){
   for(k in 1:9){
   matris[j,k] <- data[ix1[j],k]
   matris[j+100,k] <- data[ix2[j],k]
   }}
   matris <- as.data.frame(matris)


   glm_1 <- glm(Grupp ~ Datum+Tidpunkt+Belopp+Ålder+
   Poäng+Inkomst+NyKund+Kön, data=matris, family=binomial())
   summary(glm_1)


   classifier <- c(1:50)
   for(j in 1:50){
   temp <- 0
   for(k in 1:8){
   temp <- temp+(summary(glm_1)$coefficients[k+1,1])*
   data[ix3[j],k]
   }
   classifier[j] <- (summary(glm_1)$coefficients[1,1])+temp}
```

```
temp1 <- 0; temp2 <- 0
for(l in 1:50){
if(classifier[l] >= 0) {temp1 <- temp1+1} else
{temp2 <- temp2+1}
}


sim[m] <- temp1/50
}
mean(sim)
```