LUNDS UNIVERSITET

Statistiska Institutionen

Kreditklassning av aktiebolag i Sverige, en logistisk regression

Atli Ericson & Peter Ottosson

Uppsats i Statistik
10 poäng
Nivå 41-60 poäng
Februari 2007

Handledare: Mats Hagnell

ETT STORT TACK TILL

Greger Bodin och Hans Henrik Stjerna på Syna AB för all hjälp vi har fått kring deras datamaterial och verksamhet.

ABSTRACT

Predicting corporate failure is an increasingly important topic in the world of economics today. This paper, with the help of the credit ranking company Syna AB, aims to investigate a few different statistic strategies to do just that. This task is accomplished by using a data material of 250 000 Swedish companies divided into two subsamples. The first sample is used to develop the model and the second as a validation sample.

The model developing sample holds 160 000 companies divided into five different subgroups based on size and age. Binary logistic regression and probit regression were chosen for the analysis, with insolvency or not as the dependent variable. To find out which variables to include in the analysis different kinds of univariate tests are used. After the initial screening, a test of multicollinearity is also performed. Finally stepwise logistic regression is applied to lower the number of variables further.

In describing the different characteristics of each company a total of 78 variables are used as possible predictors in the model making process. Models based on logistic regression, probit regression and stepwise logistic regression alone are assessed by looking at a number of different measures. The models are compared by size, goodness of fit, significance of the parameter estimates and discriminating ability. Of the original 78 predictors about 10-25 remains in the final models created for each subgroup.

The result from the analysis implies that the most useful models are achieved through two of the three methods adopted. Stepwise logistic regression alone produces the smallest models with significant parameter estimates and high discrimination ability. Probit regression, on the other hand, manages to create models with the highest discrimination ability with the drawback of containing some insignificant parameter estimates and more variables than the previous. Unfortunately the Hosmer & Lemeshow's goodness of fit-test indicates that model fit is uncertain or poor for all models. But since other measures are satisfactory the regression models should be useful for further analysis. Because the main purpose with this paper is to make models with high predictive ability the probit regression models are chosen as the final ones.

The validation displays that in overall the final models are valid when tested against an independent sample. In two of the five subgroups the sensitivity differs when compared to the in-sample classification table. The values of the overall classification rate in the validation are 93 % for the specificity and 63 % for the sensitivity.

Innehållsförteckning

1. Inledning

Många aktörer på den ekonomiska marknaden kan dra fördel av att känna till om ett företag skulle bli oförmöget att ta hand om sina ekonomiska skyldigheter. Banker, kreditupplysningsföretag och investerare är några av de mest uppenbara intressenterna, men även företag själva kan dra nytta av sådan kännedom då tecken på ekonomisk ohälsa i det egna företaget kan fungera som en väckarklocka. Just att försöka förutse ett företags konkursbenägenhet har varit ett populärt forskningsämne de senaste 70 åren. Två ansatser har huvudsakligen använts i försöken att klassa ett företag som trolig inlämnare av konkursansökan. Dels en empirisk infallsvinkel där man ser på vilka faktorer som historiskt sett tyder på ohälsa hos ett företag, dels ett mer experimentellt förhållningssätt där man med hjälp av statistiska metoder försöker förutspå företagens benägenhet att hamna i ekonomiskt trassel.

Vilka statistiska metoder har då hittills använts? De första metoderna bestod av univariata analyser där man försökte identifiera vilka nyckeltal som bäst kunde prediktera de kommande konkurserna (Brabazon m.fl). I slutet av sextiotalet började det experimenteras med diskriminantanalys (DA) för att uppnå dessa syften. Till en början använde man endast den univariata varianten men efter hand blev multivariat diskriminantanalys (MDA) den ledande metoden. Bland tongivande studier från denna period kan Beavers univariata DA från 1966 samt Altmans MDA från 1968 nämnas (Altman, 1993). DA begränsas av flera grundläggande antaganden såsom att variabler ska vara normalfördelade, alternativt multivariat normalfördelade, vid MDA. Eftersom dessa antaganden om prediktorerna inte alltid överensstämde med verkligheten växte efter hand logistisk regression fram som ett intressant alternativ då metoden inte ställer samma snäva krav på prediktorernas fördelningar. Sedan åttiotalet har både logistisk regression och probit-regression använts flitigt i dessa sammanhang. På senare tid har även ytterligare metoder börjat användas vid kreditklassning, exempel på det är neurala nätverk där man använder sig av genetiska algoritmer som metod för variabelval.

Syna AB är ett företag inom kreditklassningsbranschen som arbetar med att ge sina uppdragsgivare information om andra aktiebolags kreditvärdighet. Frågor som de brottas med är bland annat på vilket sätt, och med vilken tillförlitlighet man kan tillmäta andra aktiebolag kreditvärdighet? Inom kreditklassningsbranschen är det av stor vikt att kunna förutspå huruvida företag är ekonomiskt pålitliga eller ej. Kreditvärdigheten hos olika företag är beroende av risken för företaget att hamna på obestånd. Med obestånd menas att vara oförmögen att betala sina skulder. Denna risk räknas ut med hjälp av data från ett stort antal indikatorer som branschen anser vara viktiga för möjligheten att förutspå obeståndsrisk. Syna kontaktade Statistiska institutionen i Lund för att hitta någon som kunde delta i arbetet med att prognostisera olika aktiebolags kreditvärdighet.

1.1. Syfte

Det huvudsakliga syftet med denna uppsats är att, utifrån ett bestämt datamaterial med aktiebolag och bakgrundsvariabler, ta fram statistiska modeller som på bästa sätt förutspår sannolikheten att hamna på obestånd. För att kunna nå detta huvudsakliga syfte måste vi också undersöka vilka statistiska metoder som lämpar sig bäst utifrån materialet och uppgiften.

2. Beskrivning av datamaterialet

Två datamaterial har använts;

-**A;** 159 966 bolag, uttaget 2004-03-30
-**B;** 90 520 bolag, uttaget 2004-03-30

Båda materialen är slumpmässigt dragna ur populationen ”Sveriges samtliga aktiebolag per den 30 mars 2004” i Synas databas. De två delarna är helt skilda från varandra och innehåller aktiebolag med uppgifter om typ, storlek, egenskaper med mera. Utav de två datamaterialen används material A för framtagande av statistiska modeller och B för validering av dessa. De aktiebolag som hamnat på obestånd inom 12 månader från det datum materialet togs ut får etiketten ”obestånd” i analysen.

2.1. Grupper

På grund av att Syna tror att en indelning av bolagen i homogenare undergrupper möjliggör en mer precis prognos, har indelningen i fyra olika grupper (**I-IV**) gjorts. Denna indelning har skapats av Syna genom att se på omsättning för varje företag, lägre gruppnummer indikerar små företag. Vidare har även ytterligare en uppdelning gjorts baserad på företagets ålder; bolag som ännu inte inkommit med något bokslut (grupp **V**). I denna del saknas även uppgift om omsättning.

I datamaterialet A skiljer sig grupperna mycket åt i storlek eftersom det finns betydligt fler små än stora företag. Exempelvis innehåller **I** ungefär 83 000 bolag medan **IV**, den minsta gruppen, består av ca: 8 000 bolag. I tabell 2.1 presenteras de fem olika indelningarna, antal observationer per grupp samt antal och andel bolag som hamnat på obestånd.

	Storlek	**Antal obs.**	**Antal obestånd**	**Andel obestånd**
I	Mycket små bolag	82 657	1531	1,85 %
II	Små bolag	38 213	689	1,80 %
III	Medelstora bolag	17 741	293	1,65 %
IV	Stora bolag	8 440	82	0,97 %
V	Unga bolag	12 915	780	6,04 %

Tabell 2.1 (Undergrupper)

2.2. Variabler

I datamaterialet finns sammanlagt 113 variabler som beskriver bolagens utformning, typ, status och egenskaper. Utav dessa variabler är det totalt 89 som mäter bolagens egenskaper och som därför går att använda som prediktorer i en analysmodell. Dessa variabler är av olika typ och är grupperade efter vilken aspekt av företagandet de vill belysa. Olika variabel-indelningar har att göra med exempelvis bolagets etableringsgrad, skötsamhet, hur det går för bolaget i termer av resultat, avkastning, likviditet och soliditet, eller egenskaper hos ledningen. Utav dessa 89 variabler tas slutligen 11 bort före analys eftersom de helt saknar

värden i det datamaterialet som används. Således återstår 78 variabler för vidare analys. Den beroende variabeln obestånd/ej obestånd (1/0) skrivs som OBEST i utskrifterna.

Vi har inte haft möjligheten att veta vad de flesta av variablerna mäter, bland annat av sekretesskäl men även för att Syna tyckte att en förutsättningslös studie skulle vara intressant. Det har för vår del inneburit att det ibland har känts lite abstrakt att inte veta exakt vilka aspekter av företagande man analyserar. Samtidigt har det förenklat analysen för oss eftersom vi inte har behövt ta hänsyn till logiskt ekonomiskt tänkande, vi har enbart behövt fokusera på variablernas typ, värden och andel saknade värden. Variablerna är av flera olika typer, vanligast är kontinuerliga och dikotoma men även andra varianter av diskreta variabler, såsom kvotvariabler, förekommer.

2.3. Saknade värden

Problemet med saknade värden är aktuellt för detta datamaterial. Andelen saknade värden kan variera kraftigt mellan variabler och gruppindelningarna, i en viss grupp kan en specifik variabel sakna många värden samtidigt som samma variabel kan ha mycket få eller inga saknade värden i någon av de andra grupperna. De medelstora bolagen, i **II** och **III**, tenderar oftare att ha saknade värden för någon av variablerna, allra störst problem med saknad data återfinns dock i grupp **V**, där de unga bolagen återfinns. Vid de datorbaserade beräkningarna stryks en observation helt ifall data saknas för någon av de aktuella variablerna. Detta gäller oavsett ifall det är den beroende eller någon av de oberoende variablerna som saknar värde. I denna uppsats hanteras problemet med saknade värden genom eliminering av de variabler som saknar många värden. Dock går det inte att undvika ett visst bortfall i flera av de skapade modellerna.

3. Teori

I detta kapitel kommer vi att gå igenom den teori som denna studie baseras på. I första delen presenteras den grundläggande teorin bakom logistisk regression samt regression via probit. Sedan går vi igenom vad man bör tänka på vid val av variabler till den logistiska modellen. Avslutningsvis diskuteras olika aspekter vid valet av slutgiltig modell samt även hur den logistiska regressionsmodellen kan tolkas. När datorbaserade beräkningar har gjorts har vi använt oss av SAS om inget annat anges, se bilaga E för användbara SAS-koder.

3.1. Logistisk regression

Med hjälp av logistisk regression ges möjligheten att diskriminera ett datamaterial mellan två eller flera grupptillhörigheter bland ett antal beroende variabler. Modellen ger den betingade sannolikheten för en observation att höra till en viss grupp, givet vissa värden på de oberoende variablerna. När utfallsvariabeln består av två grupper, i detta fall obestånd och ej obestånd, kallas den logistiska regressionen för binär.

Binär logistisk regression är besläktad med och försöker besvara samma frågor som exempelvis diskriminantanalys och multipel regression med dikotom utfallsvariabel. Fördelen med logistisk regression är att den är mer flexibel då det inte ställs samma snäva krav på

prediktorernas fördelningar. De oberoende variablerna behöver inte vara normalfördelade, ha ett linjärt samband eller ha samma varians inom grupper såsom är fallet i DA. Dessutom kan prediktorerna vara av olika typer i samma analysmaterial, kontinuerliga, diskreta eller dikotoma (Fidell & Tabachnick, 2007).

Målet med analysen är att förutspå utfallet ett eller noll. Y^* som är utfallsvariabeln i den logistiska regressionen antar sannolikheten att den beroende variabeln (y) blir ett. En ansats till att komma fram till formeln för logistisk regression är att tänka sig en underliggande latent kontinuerlig variabel Y, där utfallet blir 1 om Y > 0 och utfallet 0 om Y < 0. Y kan då beskrivas med en enkel linjär regression;

$$Y = A + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k + \varepsilon \,. \tag{3.1}$$

A är interceptet, B_j är kofficienterna och ε är slumptermen. Om man använder sig av omskrivningen;

$$u = A + \sum B_j X_{ij} \tag{3.2}$$

blir det ursprungliga uttrycket i 3.1 på formen;

$$Y = u + \varepsilon \,. \tag{3.3}$$

Sannolikheten går då att skriva som;

$$Y^* = P(Y > 0) = P(u + \varepsilon > 0) = P(\varepsilon \le u). \tag{3.4}$$

Om man antar att ε följer en logistisk fördelning och att variansen är lika med $\frac{\pi^2}{3}$ (eftersom det ger en enkel formel) blir den logistiska täthetsfunktionen;

$$\lambda(\varepsilon) = \frac{e^{\varepsilon}}{(1+e^{\varepsilon})^2} \,. \tag{3.5}$$

Med $\Lambda(u)$ som den logistiska fördelningsfunktionen går det att skriva om sannolikheten i 3.4 som;

$$P(\varepsilon \le u) = \Lambda(u) = \int_{-\infty}^{u} \frac{e^{\varepsilon}}{(1+e^{\varepsilon})^2} d\varepsilon = \frac{e^u}{1+e^u} \,. \tag{3.6}$$

Den logistiska sannolikhetsfunktionen blir då (med **i** = 1, 2,…,**n**);

$$P(y = 1 \mid x) = Y_i^* = \frac{e^u}{1+e^u} \tag{3.7}$$

Som i sin tur kan omvandlas med hjälp av inversen (den så kallade logit-transformationen) till logoddset;

$$\ln\left(\frac{Y^*}{1-Y^*}\right) = A + \sum B_j X_{ij} \tag{3.8}$$

(Long, 1997). Vilket innebär att den naturliga logaritmen av sannolikheten att klassas i en grupp delat med sannolikheten att klassas i den andra gruppen är lika med den linjära regressionsekvationen.

Till skillnad från vanlig enkel regression finns det ingen direkt intuitiv tolkning av koefficienterna i logistisk regression. Däremot kan man tolka oddskvoten vilken är lika med e^b. Oddskvoten är den multiplikativa ändringen, vid en enhets ändring hos en av prediktorerna, av oddsen att klassas i en av kategorierna (ibid). Till exempel om variabeln U1 har oddskvoten 2 innebär en förändring en enhet uppåt att sannolikheten för OBEST = 1 ökar med en faktor av 2. Uttryckt på annat sätt: sannolikheten för obestånd dubbleras. Av detta följer att en variabel med en oddskvot nära ett ger en liten påverkan på utfallet.

3.1.1 Maximum Likelihood-Skattning

För att kunna skatta koefficienterna i en logistisk regressionsekvation används metoden Maximum Likelihood-Skattning (MLE);

Den binära sannolikhetsfunktionen

$$p_i = \begin{cases} P(y_i = 1 \mid x_i) & om \quad y_i = 1 \\ 1 - P(y_i = 1 \mid x_i) & om \quad y_i = 0 \end{cases} \tag{3.9}$$

Om oberoende blir likelihoodfunktionen:

$$L(B \mid y, x) = \prod_{i=1}^{N} p_i \tag{3.10}$$

3.9 och 3.10 kombinerat ger att:

$$L(B \mid y, x) = \prod_{y=1} P(y_i = 1 \mid x_i) \prod_{y=0} (1 - P(y_i = 1 \mid x_i)) \tag{3.11}$$

eftersom,

$$P(y = 1 \mid x) = F(x, B) \tag{3.12}$$

där F är fördelningsfunktionen för logit-modellen, $\Lambda(u)$, blir den logaritmerade likelihood-funktionen:

$$\ln L(B \mid y, x) = \sum_{y=1} \ln F(x_i, B) + \sum_{y=0} \ln(1 - F(x_i, B)) \tag{3.13}$$

(Long, 1997).

Parameterskattningarna fås genom att maximera denna funktion. Med MLE följer även information om medelfelet.

3.1.2. Regression via probit

En variant av logistisk regression är probit-regression. Den producerar också sannolikheter men använder sig av standardnormalfördelningen istället för den logistiska fördelningen. Probit-funktionen blir då (med u definierat som i ekvation 3.2);

$$P(y=1 \mid X) = \Phi(u) = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2}} d\varepsilon \,. \tag{3.14}$$

MLE görs på samma sätt förutom skillnaden i fördelningarna. Koefficienterna tolkas dock på ett annat sätt, probit-koefficienten är skillnaden i standardnormalfördelningen av den beroende variabeln vid en enhets ändring av den aktuella oberoende variabeln (Fidell & Tabachnick, 2007). Probit-regression är lik den logistiska varianten och går att använda i liknande situationer, enligt Long är de i princip utbytbara (1997). Logit-variantens mer intuitivt tolkbara oddskvot ger vissa fördelar gentemot motsvarande för probit.

3.2. Val av variabler

I många fall där man använder logistisk regression så har man förkunskap om vilka variabler som är intressanta att använda, man kan därmed ana vilka variabler som analysen ska grunda sig på. I detta fall föreligger det sig dock inte på detta vis eftersom vi oftast inte vet vad variablerna står för eller vilket förklarande samband de har med utfallsvariabeln. I syfte att få ner antalet variabler till en mer hanterbar nivå kan man då använda sig av flera olika typer av analys.

3.2.1. Univariat analys

I den univariata delen av variabelanalysen undersöker man de enskilda variablernas koppling med den dikotoma utfallsvariabeln var och en för sig. Eftersom de logistiska regressions-modellerna beräknas separat på alla de fem grupperna genomförs även den univariata analysen gruppvis. För kontinuerliga variabler och kvotvariabler med ett stort antal utfall kan ett t-test för lika gruppmedelvärden ge denna koppling (Afifi & Clark, 1996). I de fallen då variablerna är dikotoma används istället ett så kallat test av två proportioner i Minitab. För kategorivariabler används ett likelihood kvot χ^2-test baserat på frekvenstabellen bestående av den beroende variabeln och variabeln i fråga (Hosmer & Lemeshow, 1989). För att inte riskera att missa användbara prediktorer väljs ofta en signifikansnivå större än 0,05, exempelvis 0,15 (Afifi & Clark, 1996).

3.2.2. Multikollinearitet

Problemen med multikollinearitet är välkända för exempelvis statistiker och ekonomer som sysslar med någon form av regressionsanalys, logistisk regression är inget undantag. Hög korrelation mellan de oberoende variablerna kan både vara av univariat och multivariat typ. Vid univariat multikollinearitet är det endast två variabler som är korrelerade, den multivariata varianten inträffar när en variabel är korrelerad med kombinationer av andra variabler. Exempel på komplikationer som kan uppstå vid multikollinearitet är att parameterskattningarnas standardfel kan höjas till oacceptabla nivåer. Vid extremt hög korrelation kan även instabila matrisinverteringar orsakas (Fidell & Tabachnick, 2007).

Via linjär regressionsanalys kan man få fram det diagnostiska värdet VIF. VIF mäter hur mycket variansen hos en skattad regressionskoefficient ökar om de oberoende variablerna är korrelerade. Det finns ingen direkt tumregel för hur höga värden på VIF som innebär problem

för modellen och olika författare förslår olika nivåer vilka sträcker sig mellan 5 och maximalt 20.

Att enbart använda sig av VIF-test, som är baserade på en linjär regressionsmodell, har dock visat sig kunna missa multikollinearitet i logistiska modeller. För att åtgärda mot heteroskedasticitet används en form av vägd minsta kvadrat-metod med viktmatrisen från ML-algoritmen från den logistiska regressionen (Davis m.fl, 1986). Se bilaga E för SAS-koder.

3.2.3. Stegvis logistisk regression

Kutner m.fl. (2004) rekommenderar stegvis logistisk regression om antal variabler överstiger 40. Varianten av stegvis logistisk regression (stepwise i SAS) som används kan ses som en kombination av framlänges stegvis logistisk regression (FSLR) och baklänges stegvis logistisk regression (BSLR). Detta eftersom den stegvisa logistiska regressionen besitter egenskaperna hos båda dessa ovannämnda varianter, då den kan ta bort variabler (BSLR) som redan inkluderats (FSLR). Vid valet av variabler i stegvis logistisk regression används Fischers scoring-metoden.

Varje steg i den stegvisa logistiska regressionen går till som följer. Enligt proceduren för FSLR beräknas först parameterskattningar av varje variabel. Därefter räknas score-χ^2-värdet ut för effekten hos varje enskild variabel. Ifall den variabel med högst score är signifikant på den valda signifikansnivån α, tas den in i modellen. Därefter testas, enligt BSLR, ifall någon av variablerna som inkluderats i modellen har ett Wald χ^2-värde (se 3.3.3 för formler) som ger ett p-värde som överstiger den valda signifikansnivån. Om så är fallet tas den aktuella variabeln bort. Signifikansnivån kan ställas in separat för både FSLR och BSLR. Proceduren upprepas tills ingen mer variabel kan tas in i modellen eller ifall variabeln som ska tas in är den enda variabeln som togs bort i föregående steg (SAS/STAT User's Guide). Se bilaga A för en SAS-utskrift som visar summeringen av variabelval i stegvis logistisk regression.

Av avgörande betydelse för antalet variabler som inkluderas i modellen är signifikansnivån, α. Grundinställningen i SAS är 0,05 men många statistiker höjer varningens finger beträffande användningen av en för låg signifikansnivå. Viktiga prediktorer, högt korrelerade till utfallet, riskerar att tryckas ut ur modellen av en enskild variabel eller en kombination av variabler (Fidell & Tabachnick, 2007). Rekommendationerna för vad som är en lämplig signifikansnivå varierar, Hosmer & Lemeshow nämner α = 0,15 eller 0,20 som rimliga nivåer (1989). Ifall syftet med den stegvisa logistiska regressionen är att få med många variabler, för att därmed få en mer komplett bild av möjliga samband mellan variablerna, kan en högre signifikansnivå användas, exempelvis α =0,25 (ibid).

3.3. Modellval

Efter variabelreduktionen, då man lyckats få fram de variabler som har stor betydelse för modellbygget, återstår det avslutande arbetet med att välja den modell som bäst passar de syften man har. Hur kan man bedöma vilken modell som är bäst? I denna del kommer vi att redovisa de viktigaste aspekterna för de slutgiltiga valen av variabler samt de tester och mått som behövs för att visa på det. I bilaga B finns en komplett utskrift av en körning med logistisk regression.

3.3.1. Förenkling av modellen

Vid arbetet med att ta fram en regressionsmodell strävar man ofta efter att få ner antalet variabler. Syftet med att minimera antal variabler i en modell, förutom att den blir enklare att arbeta med, är att ju enklare modell desto mer numerisk stabil blir den. En modell med flera variabler innebär även ökade medelfel för parameterskattningarna (Hosmer & Lemeshow, 1989). Dessutom är en mindre modell inte lika beroende av datamaterialet då färre variabler generellt innebär mindre risk för saknade värden.

3.3.2. Passar modellen datamaterialet?

Det finns olika sätt att undersöka ifall den beräknade regressionsmodellen är anpassad till datamaterialet. Ett test som Kutner m.fl. (2004) rekommenderar vid data utan replikat är Hosmer & Lemeshow's anpassningstest (HL-test), där observationerna delas upp i tio delar baserade på de beräknade sannolikheterna. Därefter appliceras ett χ^2-test beräknat på observerade och förväntade värden på utfallsvariabeln. Ett p-värde räknas fram från χ^2-fördelningen med åtta frihetsgrader för att testa hur väl den logistiska modellen överensstämmer med datamaterialet. Ett icke signifikant värde, större än 0,05, visar på god anpassning (Hosmer & Lemeshow, 1989).

Det förekommer dock diskussioner om restriktioner till ovanstående test. Fidell & Tabachnick skriver att vid stora stickprov kan även små skillnader mellan observerade och förväntade värden bli signifikanta (2007). Restriktioner beträffande cellfrekvenser är ett omstritt område och rekommendationerna skiftar mellan olika författare. Vissa statistiker menar att ingen förväntad cellfrekvens bör vara under 5 (Hosmer & Lemeshow, 1989). Själva är Hosmer & Lemeshow mer liberala och anser att cellfrekvenser visst kan vara under 5 och till och med under 1 i vissa fall. Ytterligare rekommendationer innebär att ingen förväntad cellfrekvens får ha ett värde under 1 samt att endast en cell får ha ett värde under 5 (Garson, Statnotes: Topics in Multivariate Analysis).

3.3.3. De enskilda parameterskattningarnas signifikans

Det finns två test som är aktuella som signifikansmått för de enskilda parametrarna. Det första och enklaste testet är Wald χ^2, vilket beräknas med kvoten av parameterskattningen i kvadrat genom dess medelfel i kvadrat. Hypoteserna ser ut som följande;

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

Med testfunktionen:

$$z^* = \frac{b_k}{s(b_k)} \qquad (3.15)$$

och om $z^{*2} > \chi^2(1)$ förkastas nollhypotesen. Värdet α är vald signifikansnivå (Kutner m.fl, 2004).

Det andra testet som kan användas för att pröva signifikansen hos en eller flera variabler, och som dessutom kan användas för att jämföra olika modeller med varandra, heter likelihood kvot-test av skillnad i χ^2. I praktiken testas variabler genom att man jämför en modell där variablerna är inkluderade med en modell där variablerna tagits bort. Därefter beräknas de olika modellernas -2 * log-likelihood-värden som sedan jämförs med χ^2-fördelningen enligt:

$$\chi^2 = \left[-2\log L(M) - \left(-2\log L(S)\right) \right]. \tag{3.16}$$

Frihetsgraderna blir antal parametrar i den större (S) modellen minus antalet i den mindre (M) (ibid, 2007). Till exempel vid en jämförelse mellan två modeller där den ena har en extra parameter, vilket ger en frihetsgrad, blir det kritiska värdet för χ^2-fördelningen 3,84 på 5 % -nivån.

3.3.4. Mått på klassificering

Eftersom det främsta syftet med denna uppsats är att försöka få fram modeller med så hög grad av prediktion som möjligt så är det av intresse att redovisa på vilka sätt dessa modellers klassificeringsförmåga kan beskrivas.

För att undersöka den prediktiva kraften i modellen kan man i SAS få fram flera mått baserade på rangordnade parvisa observationer. De tas fram genom att först para ihop de observationer där utfallen skiljer sig åt så att sammanlagt t par skapas. Parens beräknade sannolikheter rangordnas och jämförs sedan med observationernas faktiska utfall. I de fall där utfallet 1, det vill säga obestånd, också innehar den högsta sannolikheten benämns paret överensstämmande (n_c), annars motstridigt (n_d). Om paret varken är motstridigt eller överensstämmande benämns det som lika, t - n_c - n_d par är lika. Måttet som används i denna undersökning är

$$Somers'\,D = \left(n_c - n_d\right)/t \tag{3.17}$$

(SAS/STAT User's Guide). Somers' D går också under beteckningen AR (accuracy ratio), en omskrivning som anammas i denna uppsats. Beteckningen AR är rådande inom kredit-klassningsbranschen och används bland annat i statistikprogrammet Minitab. För beviset att Somers' D = AR, se Hamerle m fl. (2003). Måttet AR kan även ses som modellens förmåga att diskriminera mellan de två grupperna obestånd eller ej obestånd.

Med kommandot CTABLE i SAS fås klassificeringstabellen för körningen. Varje företag i stickprovet klassas utifrån den logistiska regressionen och antal korrekta klassificeringar för båda grupperna skrivs ut. Detta beräknas för flera olika värden på den avgränsande sannolikheten. I tabellen refererar Event till utfallet 1, de två kolumnerna Correct till antal rätt klassade observationer för de två grupperna, Incorrect Event är antal faktiska ”ej obestånd” som klassas fel (typ 1-fel eller ”falskt alarm”) och Incorrect Nonevent antal faktiska obestånd som klassas fel (typ 2-fel eller ”miss”). Sensitiviteten är andelen obestånd som klassas rätt och specificiteten är andelen rätt skattade ”ej obestånd”. Därmed återstår frågan vilken sannolikhet man ska välja som gräns för sin klassificering.

Kutner m.fl. (2004) har en diskussion kring hur skiljevärdet för ens klassificering väljs. De faktorer som spelar in är hur datamaterialet ser ut och huruvida kostnaden för typ 1 och typ 2-

fel är lika. Eftersom andelen obestånd i datamaterialet i snitt är ca 2 % blir den tidigare sannolikheten ganska liten vilket tyder på att sätta gränsen för klassificeringen till 0,5 inte är helt rimlig.

Kostnaderna för de två typerna av felklassning är i det verkliga fallet knappast samma. En bank som behandlar en låneansökan från ett företag skulle förmodligen hellre bedöma ett friskt företag som konkursmässigt än tvärt om. Eftersom den här undersökningen är gjord i ett allmänt syfte har vi valt att bortse från skillnader i felens kostnader. Om man inte har information om felens kostnader rekommenderar Kutner att använda sig av de tidigare sannolikheterna och försöka maximera andelen rätt klassade företag. Detta kan göras genom att studera klassificeringstabellen och försöka sätta avgränsningen utifrån den (2004). En fara är att tabellen grundar sig på de data som den framtagna modellen baseras på. Därför kan klassificeringstabellens korsvalidering av modellen ge optimistiska mått på förmågan att klassificera mellan grupperna. I SAS används en approximation för att efterlikna resultatet av en så kallad Leave one out-korsvalidering (Schlotzhauer). Den typen av korsvalidering innebär att en observation används till validering medan resten av datamaterialet används till skattning av modellen. Detta upprepas sedan för alla återstående observationer.

3.3.5 Begränsningar hos den logistiska regressionen

För att kunna skapa bästa möjliga logistiska regressionsmodell går det att undersöka ytterligare en faktor som kan påverka resultatet. Logistisk regression kräver inte det linjära förhållandet mellan variablerna som enkel regression gör, däremot är linjäritet mellan prediktorerna och logit-transformationen av den beroende variabeln ett av de grundläggande antagandena. En metod att testa detta antagande är att lägga till ett antal variabler till den logistiska modellen. De nya variablerna skapas som ett samspel mellan varje oberoende variabel och dess naturliga logaritm. Vid signifikans av någon eller några av dessa samspel tyder det på att antagandet om linjäritet inte är uppfyllt. Fidell & Tabachnick rekommenderar transformation av de variabler som inte uppfyller kraven (2007). Eftersom den här undersökningen är baserad på ett så stort antal variabler har vi valt att inte utföra det här testet.

4. Resultat

I detta kapitel genomförs och tolkas de statistiska körningar som tagits upp i kapitel tre. Upplägget i detta kapitel kommer i stort att följa den ordning som kapitel tre har; först väljs lämpliga variabler ut för att sedan användas i de slutgiltiga logistiska regressionsmodellerna.

4.1. Variabelval

4.1.1. Variabelreduktion efter univariat analys och multikollinearitetstest

Den första delen av analysen går ut på att ta bort de oberoende variabler som antingen inte förklarar något av variationen i utfallsvariabeln eller som uppvisar hög multikollinearitet sinsemellan. Vilka variabler som väljs bort visar sig skifta mycket mellan de olika grupperna,

en skillnad som finns både beträffande brist på förklaring och beträffande hög multikollinearitet. Någon genomgående trend för borttagna variabler mellan grupperna går således inte att spåra, dock framgår det av resultatet att färre variabler kan tas bort i **II** och **III** än i övriga grupper. I den univariata analysen har signifikansnivån 0,15 använts som gräns. I testet för multikollinearitet har vi bedömt VIF-värden runt 10 som allvarlig korrelation. Tabell 4.1 visar vilka variabler som väljs bort efter dessa två inledande analyser.

	Betalning	Etablering	Ledning	Utveckling	Övriga	Antal/total
I			l16, l20, l29, l31	Alla utom u3	a1	29/78
II	b4	e2, e4	l1, l20, l21, l28, l30, l32, l33	u1, u2, u4, u5, u11, u12, u13, u14, u15, u16, u17, u19, u22		22/78
III	b2		l4, l14, l16, l19, l20, l21, l26, l27, l31	u2, u3, u4, u5, u8, u9, u10, u11, u14, u15		20/78
IV	b2, b3, b4	e6, e7, e8, e9, e10, e11, e12,	l4, l5, l9, l12, l13, l14, l16, l21, l26, l27, l31, l33	u1, u2, u6, u7, u11, u12, u14, u15, u16, u17	d1	33/78
V	b4	e1, e6, e10	l26, l31, l32	Alla	ek1, a1, d4	35/78

Tabell 4.1 (Borttagna variabler efter univariat analys och multikollinearitetstest)

4.1.2. Variabelreduktion genom stegvis logistisk regression

Den stegvisa logistiska regressionen genomförs därefter på återstående variabler. Olika α-värden, varierande mellan 0,2 och 0,3, har både testats och använts för de olika grupperna. Samma signifikansnivå har använts för både FSLR och BSLR. De variabler som väljs bort i tabell 4.2 faller ur modellerna på signifikansnivån $\alpha = 0{,}25$.

	Betalning	Etablering	Ledning	Utveckling	Övriga	Antal/total
I	b6	e2, e6, e7, e9, e10, e12	l4, l5, l12, l14, l19, l22, l23, l26		ek1, ek2, ek4	18/49
II	b1, b2, b6	e7, e8, e10, e11, e12	l4, l5, l9, l11, l12, l13, l14, l15, l18, l26, l27, l31, l32	e6, e7, e8, e9, e10, e21, e23, e24, e25	d1, d2, ek1, ek2, ek4	35/56
III	b4	e1, e2, e7, e8, e9, e10, e11, e12	l1, l10, l11, l12, l13, l15, l23, l28, l29, l32, l33	u1, u6, u7, u13, u16, u17, u19, u20, u21, u22, u23, u24, u25	d1, d2, ek1, ek2, ek4	38/58
IV	b1, b5, b6	e2, e4	l10, l11, l15, l18, l19, l20, l22, l23, l28, l30, l32	u3, u4, u5, u8, u9, u10, u13, u14, u15, u21, u22, u23, u24, u25	d2, ek1, ek2, ek4	34/45
V	b6	e7, e11, 12	l4, l9, l11, l15, l16, l19, l20, l21, l22, l28, l30, l33		ek2, ek3, ek4	19/43

Tabell 4.2 (Borttagna variabler efter stegvis logistisk regression)

4.2. Val av regressionsmodeller

Målet vid skapandet av de olika slutgiltiga regressionsmodellerna är att förenkla regressionsmodellen för de återstående variablerna och att samtidigt behålla eller höja modellens diskrimineringsförmåga. Med diskrimineringsförmåga åsyftas måttet AR. De variabler som tas ur modellen har antingen höga medelfel på parameterskattningarna eller låg inverkan på utfallsvariabeln. För att undersöka lämpligheten i att ta bort en variabel kontrolleras de enskilda variablernas inverkan på modellen med hjälp av likelihood kvot-testet enligt avsnitt 3.3.3.

Vid modifieringen av de slutgiltiga modellerna finns ofta en motsättning mellan antal variabler i modellen och modellens diskrimineringsförmåga. Ofta når modellen sin högsta diskrimineringsförmåga när ett antal variabler med ickesignifikanta p-värden i Wald-testet finns med. Att ta bort en av dessa variabler är möjligt när χ^2-värdet i likelihood kvot-testet understiger 3,84. Tyvärr leder dock elimineringen av en variabel ofta även till att diskrimineringsförmåga försvinner. Vad ska man då prioritera högst? Liten modellstorlek och signifikanta p-värden i Wald-testet eller en mindre strikt syn på modellstorlek och parameterskattningarnas signifikans ifall det leder till ökad diskrimineringsförmåga. Något givet svar på denna avvägningsfråga finns knappast utan är snarare beroende av vilket ändamål man har. Vid de regressionsmodeller som är aktuella i detta fall är förlusterna i diskrimineringsförmåga ibland ganska små i förhållande till hur pass mycket enklare modellen kan bli. I dessa fall är det naturligtvis ännu mer tveksamt om man ska sätta diskrimineringsförmågan främst.

Ett annat sätt att mäta lämpligheten hos olika modeller är att se på de värden som fås med Hosmer & Lemehow's anpassnings-test. Testresultaten vid körningarna på detta datamaterial ger kraftigt signifikanta p-värden utom i grupp **III** och till viss del **IV**, något som indikerar att de flesta modellerna inte är anpassade till datamaterialet. Vad detta beror på kan vara intressant att diskutera. Värt att notera är tendensen att grupperna med större antal observationer, **I** och **II**, har sämre värden i detta test. Det stora antalet observationer i dessa grupper kan eventuellt vara en förklaring till de låga p-värden i testet då antalet observationer kan påverka signifikansnivån. En närmare undersökning av cellfrekvenserna för alla grupper visar att **III** och **IV**, som har bäst värden i testet, har låga eller mycket låga förväntade cellfrekvenser för de lägsta sannolikheterna. Eftersom de två grupperna inte uppfyller restriktionerna går det därför inte med säkerhet att dra slutsatsen att modellerna är anpassade till datamaterialet utifrån HL-testet. Övriga grupper har förväntade cellfrekvenser som med marginal överstiger de mest restriktiva tolkningarna. Sammantaget ger HL-testen dåliga eller osäkra resultat för samtliga modeller. Enligt Hosmer & Lemeshow är det trots detta relevant att fortsätta utvärdera resultatet ifall modellerna visar på god diskriminering (2000).

Vid modifieringen av de slutgiltiga logistiska regressionsmodellerna visar sig länkfunktionen probit vara bättre än logit beträffande diskrimineringsförmåga i kombination med modellstorlek och parameterskattningarnas signifikans. Tabell 4.3 visar skillnaden i diskrimineringsförmåga, beskrivet med hjälp av måttet AR, mellan olika varianter av modellbygge.

Länk-funktion →	Logit	Logit	Probit
	Enbart stegvis α =0,05	Stegvis plus modellval	Stegvis plus modellval
I	0,834	0,832	0,837
II	0,760	0,765	0,779
III	0,701	0,699	0,711
IV	0,804	0,802	0,808
V	0,802	0,805	0,805

Tabell 4.3 (Accuracy Ratio för de olika modellerna)

I tabellen ser man tydligt att den högsta diskrimineringsförmågan fås när man använder sig av probit-analys vid modifieringen av den slutgiltiga modellen. Störst skillnad mellan de olika typerna av modeller nås i **II** och **III**, skillnaden i övriga grupper är inte särskilt stora. Där skillnaden är minst, i **I**, **IV** och **V**, aktualiseras därför frågan om man ska sätta diskrimineringsförmågan eller en enklare modell med signifikanta parametrar i fokus vid modellval.

I den första gruppen har den probit-baserade modellen 27 variabler varav 4 koefficienter är icke-signifikanta på 5 % -nivån och den högsta p = 0,1832. Modellen där enbart stegvis logistisk regression använts har 24 variabler och alla koefficienter är naturligtvis signifikanta på 5 % -nivån. HL-testet visar dock att den probit-baserade modellen har betydligt bättre χ^2-värde (35,5 mot 62,5). I **IV** har den probit-baserade modellen 11 variabler mot 10 för den modell som enbart tagits fram men hjälp av stegvis logistisk regression. En koefficient är icke-signifikant på 5 % -nivån hos probit-modellen (p = 0,111). Även här visar sig den probit vara överlägsen i HL-testet (χ^2-värde = 15,5 mot 20,7). I **V** har modellen framtagen med enbart stegvis logistisk regression 16 variabler mot 18 för probit, varav en koefficient är icke-signifikant (p = 0,0733). Skillnaden i HL-testet är små (χ^2 = 25,4 för enbart stegvis och 23,4 för probit). Modellen där logit använts vid valet av slutgiltig regressionsmodell väljs bort då den innehåller fler variabler än de övriga och dessutom har flera icke-signifikanta parameterskattningar på 5 % -nivån.

Värt att notera i grupp **III**, där skillnaden i diskrimineringsförmåga mellan modellen baserad på probit och modellen framtagen med enbart stegvis är påtaglig, är att probit-modellen innehåller 17 variabler mot 12 för den enbart stegvisa. Utav dessa 17 variabler är dessutom 7 koefficienter icke-signifikanta på 5 % -nivån. Skillnaderna i HL-testen är marginella.

Vid val av modeller undersöks även ifall den slutgiltiga modellen innehåller variabler som saknar många värden. Detta visade sig dock inte vara något större problem då andelen saknade värden som högst uppgick till 15 procent (**III**, probit).

4.3. Slutgiltiga modeller och dess egenskaper

De modeller som vi bedömer vara bäst utifrån syftet med uppsatsen är alla skapade med probit-regression. Huruvida man borde välja någon av de enklare modellerna där enbart stegvis logistisk regression använts kan diskuteras, men då syftet med uppsatsen främst är att skapa modeller med hög diskrimineringsförmåga så anser vi att dessa val är rimliga.

Utskrifter för samtliga modellers parameterskattningar, mått på diskrimineringsförmåga, HL-test och klassificeringstabeller finns i bilaga C.

	AR	**Antal variabler**	**Antal icke-signifikanta koefficienter**
I	0,837	27	4
II	0,779	21	2
III	0,711	17	7
IV	0,808	11	1
V	0,805	18	1

Tabell 4.4 (Sammanfattande egenskaper för de bästa modellerna)

4.4. Valideringsresultat

För att slutgiltigt undersöka hur väl modellerna fungerar testas ekvationerna på ett nytt material, datamaterial B. Detta utförs genom att man klassificerar de nya observationerna utifrån de slutgiltigt valda modellerna. För att kunna klassificera mellan de båda utfallen krävs det att ett skiljevärde väljs, de värden som verkar klassificera bäst enligt resonemanget i 3.4.3 är antingen 2 % eller 4 %. I bilaga D återfinns korstabeller för de olika klassificeringarna.

Varje tabell kan liknas vid en rad av klassificeringstabellen som redogörs för i 3.4.3. Raderna är de faktiska utfallen och kolumnerna de skattade sådana, under frekvenserna står andelarna för skattningarna. Till exempel i grupp **I**, med 0,04 som skiljevärde, är andelen rätt skattade obestånd (sensitiviteten) 65,83 %. Motsvarande värde för den ursprungliga klassificerings-tabellen gjord på material A är 67,5 %.

Specificiteten håller sig på nivå runt 90 % vilket får anses som ett bra resultat om än inte helt oväntat på grund av den höga andelen *ej obestånd*. Sensitiviteten 40 % för III respektive 29,41 % för IV ser dock inte så imponerande ut. En jämförelse med motsvarande siffror i analysmaterialet, 57, 4 % respektive 64,5 %, indikerar att de modellernas prediktiva förmåga möjligtvis inte är så bra. Dock är antal obestånd i valideringsmaterialet för **IV** bara 34 stycken vilket möjligen kan vara ett för litet antal för en god validering

Sammanfattningsvis kan vi konstatera att modellernas förmåga att klassificera nya observationer i de flesta fallen är nära det resultat som de tidigare klassificeringstabellerna indikerat, samt att modellernas diskrimineringsförmåga klart överstiger slumpen. Som ett sammanfattande mått är specificiteten för hela material B 93,08 % och sensitiviteten 63,25 %. Detta beräknas som totala antalet rätt klassificerade genom totala antalet observationer. I tabell 4.5 redovisas valda skiljevärden och resultaten för sensitivitet och specificitet.

	Skiljevärde på slh.	Sensitivitet i %	Specificitet i %
I	0,04	65,83	95,57
II	0,02	59,17	90,14
III	0,02	40,00	92,50
IV	0,02	29,41	91,32
V	0,04	73,37	87,45

Tabell 4.5 (Resultat av valideringen)

5. Slutsats

En uppenbar slutsats som kan dras är att probit fungerade bättre än logit när det gäller prognostiseringen av obestånd i det använda datamaterialet. Stegvis logistisk regression visade sig vara en effektiv metod att välja ut variabler på. Eventuellt kan det bero på att metoden lämpar sig väl vid stora datamaterial med många variabler, något som litteraturen på ämnet antyder. Stegvis logistisk regression fungerade inte bara väl beträffande valen av variabler utan även vid framtagandet av modeller och parameterskattningar. De modeller som skapats av enbart stegvis logistisk regression har god diskrimineringsförmåga och kan mycket väl tänkas användas om man är ute efter enklare modeller. Eftersom uppsatsens huvudsakliga syfte är att ta fram statistiska modeller som på bästa sätt förutspår sannolikheten att hamna på obestånd väljs probit-modellerna till slutgiltiga modeller då de uppvisar bäst diskrimineringsförmåga.

HL-testerna visade generellt att de skapade modellerna inte är särskilt väl anpassade till datamaterialet då p-värdena för flera grupper blev mycket låga, vad detta beror på kan inte fastställas med säkerhet. Då modellernas diskrimineringsförmåga får anses vara god samt att valideringen visar att de slutgiltiga modellerna generellt klassificerar väl på ett nytt datamaterial bedöms modellerna trots allt vara bra.

Referenser

Afifi, A.A & Clark, V. 1996. *Computer-aided Multivariate Analysi*s. Chapman & Hall. London

Altman, Edward I, 1993. *Corporate financial distress and bankruptcy*. John Wiley & Sons, Inc. New York

Brabazon, Mathews, O'Neill & Ryan, Grammatical Evolution and Corporate Failure Prediction, *GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference*, (2002), 1011-1018.

Fidell, Linda S & Tabachnick, Barbara G. 2007. *Using Multivariate Statistics*. Pearson Education Inc. Boston

Davis, C. E, Hyde, J. E, Bangdiwala, S. I & Nelson , J. J. 1986. An Example of Dependencies Among Variables in a Conditional Logistic Regression, *Modern Statistical Methods in Chronic Disease Epidemiology*, John Wiley & Sons Inc. New York

Hosmer, David W. & Lemeshow, Stanley. 1989. *Applied Logistic Regression*, John Wiley & Sons Inc. New York

Hosmer, David W. & Lemeshow, Stanley. 2000. *Applied Logistic Regression*; *second edition*, John Wiley & Sons Inc. New York

Kutner, Nachtsheim, Neter, 2004. *Applied Linear Regression Models*, McGraw-Hill/Irwin, New York

Long, J. Scott, 1997. *Regression Models for Categorical and Limited Dependent Variables*, SAGE Publications, London

Internetreferenser

Garson, David G, *Statnotes: Topics in Multivariate Analysis*, http://www2.chass.ncsu.edu/garson/pa765/logistic.htm, 2007-01-12

SAS/STAT User's Guide, Chapter 39, The Logistic Procedure, http://www.math.wpi.edu/saspdf/stat/, 2007-01-12

Schlotzhauer, David C. *Some Issues in Using PROC LOGISTIC for Binary Logistic Regression*, http://www.ats.ucla.edu/stat/sas/library/ts274.pdf, 2007-01-12

Hamerle, Alfred, Rauhmeier, Robert, Rösch, Daniel, 2003, *Uses and Misuses of Measures for Credit Rating Accuracy*, http://www.defaultrisk.com/pp_test_25.htm, 2007-01-12

Bilaga A

Summering av en stegvis logistisk regression, grupp 3

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	l6		1	1	2392.7640	.	<.0001	RLEDC30
2	b10		1	2	927.8105	.	<.0001	RBETA15
3	u18		1	3	96.4697	.	<.0001	RUTVD20
4	b3		1	4	44.6079	.	<.0001	RBETA30
5	l8		1	5	36.1231	.	<.0001	RLEDD10
6	e6		1	6	26.9060	.	<.0001	RETAF10
7	a1		1	7	26.6894	.	<.0001	RATSNIK
8	l28		1	8	22.4429	.	<.0001	RLEDC00
9	RANTOMFR		1	9	15.5722	.	<.0001	RANTOMFR
10	d4		1	10	13.6107	.	0.0002	RDIVA30NUM
11	b5		1	11	13.1106	.	0.0003	RBETB10
12	l23		1	12	9.6954	.	0.0018	RLEDG10
13	ek3		1	13	8.8351	.	0.0030	REKOA25
14	l10		1	14	7.0143	.	0.0081	RLEDD30
15	e4		1	15	3.9250	.	0.0476	RETAD10
16	e9		1	16	2.3424	.	0.1259	RETAF18
17	l18		1	17	2.4638	.	0.1165	RLEDF10
18	l16		1	18	2.5631	.	0.1094	RLEDE20
19	u20		1	19	1.8372	.	0.1753	RUTVD40
20	l22		1	20	1.6044	.	0.2053	RLEDF50
21		l18	1	19	.	1.5516	0.2129	RLEDF10
22	u3		1	20	1.5066	.	0.2197	RUTVA20
23		u3	1	19	.	1.5051	0.2199	RUTVA20

Bilaga B

Exempel på enkel logistisk regression, fullständig utskrift, grupp 4

The LOGISTIC Procedure

Model Information

Data Set	SASBIBL.RATZ4	
Response Variable	ROBEST3NY	ROBEST3NY
Number of Response Levels	2	
Number of Observations	8198	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Response Profile

Ordered Value	ROBEST3NY	Total Frequency
1	1	76
2	0	8122

Probability modeled is ROBEST3NY=1.

NOTE: 241 observations were deleted due to missing values for the response or explanatory variables.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	864.792	664.521
SC	871.804	748.661
-2 Log L	862.792	640.521

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	222.2710	11	<.0001
Score	592.0704	11	<.0001
Wald	180.0850	11	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4141	0.4318	31.2628	<.0001
b10	1	0.0841	0.0363	5.3626	0.0206
e1	1	-0.00179	0.000869	4.2296	0.0397
l1	1	-0.1925	0.0789	5.9500	0.0147
l6	1	3.7080	0.8183	20.5329	<.0001
l8	1	0.0181	0.00510	12.5510	0.0004
l29	1	-0.00138	0.000439	9.9083	0.0016
u18	1	-0.0538	0.0105	26.2122	<.0001
RANTOMFR	1	2.7636	0.7765	12.6682	0.0004
ek3	1	1.0611	0.3663	8.3921	0.0038
d4	1	1.7749	0.4171	18.1077	<.0001
a1	1	0.0785	0.0538	2.1299	0.1445

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
b10	1.088	1.013	1.168
e1	0.998	0.997	1.000
l1	0.825	0.707	0.963
l6	40.772	8.200	202.725
l8	1.018	1.008	1.028
l29	0.999	0.998	0.999
u18	0.948	0.928	0.967
RANTOMFR	15.857	3.462	72.633
ek3	2.889	1.409	5.924
d4	5.900	2.605	13.362
a1	1.082	0.973	1.202

Association of Predicted Probabilities and Observed Responses

Percent Concordant	88.6	Somers' D	0.802
Percent Discordant	8.4	Gamma	0.827
Percent Tied	3.0	Tau-a	0.015
Pairs	617272	c	0.901

The LOGISTIC Procedure

Partition for the Hosmer and Lemeshow Test

Group	Total	ROBEST3NY = 1 Observed	ROBEST3NY = 1 Expected	ROBEST3NY = 0 Observed	ROBEST3NY = 0 Expected
1	731	1	0.07	730	730.93
2	755	0	0.22	755	754.78
3	755	1	0.49	754	754.51
4	858	1	1.01	857	856.99
5	816	0	1.63	816	814.37
6	803	0	2.48	803	800.52
7	803	2	3.78	801	799.22
8	828	4	6.06	824	821.94
9	824	7	9.90	817	814.10
10	1025	60	50.02	965	974.98

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
21.1399	8	0.0068

NOTE: In calculating the Expected values, predicted probabilities less than 0.0001 and greater than 0.9999 were changed to 0.0001 and 0.9999 respectively.

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	76	0	8122	0	0.9	100.0	0.0	99.1	.
0.020	45	7458	664	31	91.5	59.2	91.8	93.7	0.4
0.040	25	7900	222	51	96.7	32.9	97.3	89.9	0.6
0.060	25	7991	131	51	97.8	32.9	98.4	84.0	0.6
0.080	22	8032	90	54	98.2	28.9	98.9	80.4	0.7
0.100	20	8047	75	56	98.4	26.3	99.1	78.9	0.7
0.120	17	8065	57	59	98.6	22.4	99.3	77.0	0.7
0.140	14	8081	41	62	98.7	18.4	99.5	74.5	0.8
0.160	13	8089	33	63	98.8	17.1	99.6	71.7	0.8
0.180	11	8093	29	65	98.9	14.5	99.6	72.5	0.8
0.200	10	8097	25	66	98.9	13.2	99.7	71.4	0.8
0.220	9	8097	25	67	98.9	11.8	99.7	73.5	0.8
0.240	9	8100	22	67	98.9	11.8	99.7	71.0	0.8
0.260	9	8102	20	67	98.9	11.8	99.8	69.0	0.8
0.280	8	8103	19	68	98.9	10.5	99.8	70.4	0.8
0.300	7	8104	18	69	98.9	9.2	99.8	72.0	0.8
0.320	7	8107	15	69	99.0	9.2	99.8	68.2	0.8
0.340	6	8107	15	70	99.0	7.9	99.8	71.4	0.9

The LOGISTIC Procedure

Classification Table

Prob Level	Correct Event	Correct Non-Event	Incorrect Event	Incorrect Non-Event	Percentages Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.360	6	8109	13	70	99.0	7.9	99.8	68.4	0.9
0.380	6	8111	11	70	99.0	7.9	99.9	64.7	0.9
0.400	6	8114	8	70	99.0	7.9	99.9	57.1	0.9
0.420	6	8114	8	70	99.0	7.9	99.9	57.1	0.9
0.440	6	8114	8	70	99.0	7.9	99.9	57.1	0.9
0.460	6	8115	7	70	99.1	7.9	99.9	53.8	0.9
0.480	5	8115	7	71	99.0	6.6	99.9	58.3	0.9
0.500	5	8116	6	71	99.1	6.6	99.9	54.5	0.9
0.520	5	8116	6	71	99.1	6.6	99.9	54.5	0.9
0.540	5	8116	6	71	99.1	6.6	99.9	54.5	0.9
0.560	5	8116	6	71	99.1	6.6	99.9	54.5	0.9
0.580	4	8116	6	72	99.0	5.3	99.9	60.0	0.9
0.600	4	8117	5	72	99.1	5.3	99.9	55.6	0.9
0.620	4	8117	5	72	99.1	5.3	99.9	55.6	0.9
0.640	4	8117	5	72	99.1	5.3	99.9	55.6	0.9
0.660	4	8117	5	72	99.1	5.3	99.9	55.6	0.9
0.680	4	8118	4	72	99.1	5.3	100.0	50.0	0.9
0.700	4	8119	3	72	99.1	5.3	100.0	42.9	0.9
0.720	4	8120	2	72	99.1	5.3	100.0	33.3	0.9
0.740	4	8120	2	72	99.1	5.3	100.0	33.3	0.9
0.760	3	8120	2	73	99.1	3.9	100.0	40.0	0.9
0.780	3	8120	2	73	99.1	3.9	100.0	40.0	0.9
0.800	3	8120	2	73	99.1	3.9	100.0	40.0	0.9
0.820	3	8121	1	73	99.1	3.9	100.0	25.0	0.9
0.840	2	8122	0	74	99.1	2.6	100.0	0.0	0.9
0.860	2	8122	0	74	99.1	2.6	100.0	0.0	0.9
0.880	2	8122	0	74	99.1	2.6	100.0	0.0	0.9
0.900	2	8122	0	74	99.1	2.6	100.0	0.0	0.9
0.920	2	8122	0	74	99.1	2.6	100.0	0.0	0.9
0.940	2	8122	0	74	99.1	2.6	100.0	0.0	0.9
0.960	1	8122	0	75	99.1	1.3	100.0	0.0	0.9
0.980	1	8122	0	75	99.1	1.3	100.0	0.0	0.9
1.000	0	8122	0	76	99.1	0.0	100.0	.	0.9

Bilaga C

Beskrivning av de viktigaste egenskaperna hos de slutgiltiga modellerna, probit-regression, grupp **I - V**

Grupp I

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.1482	0.0634	1147.7895	<.0001
b2	1	0.00944	0.00251	14.1268	0.0002
b3	1	0.1458	0.0606	5.7912	0.0161
b5	1	0.0791	0.0151	27.2824	<.0001
b10	1	0.2635	0.0130	407.9107	<.0001
e1	1	-0.00043	0.000138	9.6882	0.0019
e8	1	0.0983	0.0464	4.4866	0.0342
e11	1	0.1168	0.0638	3.3495	0.0672
l1	1	-0.0457	0.0186	6.0327	0.0140
l6	1	0.9173	0.0825	123.6219	<.0001
l8	1	0.00296	0.000516	32.8507	<.0001
l10	1	0.00293	0.000610	22.9897	<.0001
l11	1	0.00114	0.000683	2.7956	0.0945
l13	1	0.00199	0.000776	6.5863	0.0103
l15	1	-0.00273	0.00150	3.3215	0.0684
l18	1	-0.00032	0.000062	26.7371	<.0001
l21	1	-0.00068	0.000156	19.2980	<.0001
l26	1	-0.1984	0.0939	4.4663	0.0346
l27	1	0.00872	0.00266	10.7387	0.0010
l28	1	0.5298	0.0663	63.8365	<.0001
l32	1	-0.00342	0.000481	50.7113	<.0001
l33	1	0.00180	0.000608	8.7852	0.0030
u3	1	0.000037	0.000028	1.7714	0.1832
RANTOMFR	1	2.2407	0.4261	27.6577	<.0001
ek3	1	0.3326	0.0235	199.8903	<.0001
d1	1	0.4793	0.0372	166.0704	<.0001
d2	1	0.1415	0.0632	5.0185	0.0251
d4	1	0.3359	0.0449	55.9404	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	90.9	Somers' D	0.837
Percent Discordant	7.1	Gamma	0.854
Percent Tied	2.0	Tau-a	0.030
Pairs	124200844	c	0.919

The LOGISTIC Procedure

Partition for the Hosmer and Lemeshow Test

		OBEST = 1		OBEST = 0	
Group	Total	Observed	Expected	Observed	Expected
1	7793	8	7.10	7785	7785.90
2	8720	3	15.43	8717	8704.57
3	7849	7	18.04	7842	7830.96
4	8636	13	24.93	8623	8611.07
5	8036	20	29.91	8016	8006.09
6	8351	35	41.43	8316	8309.57
7	8220	52	55.48	8168	8164.52
8	8326	80	79.97	8246	8246.03
9	8273	152	132.70	8121	8140.30
10	8451	1161	1090.25	7290	7360.75

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
35.2856	8	<.0001

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	1531	0	81124	0	1.9	100.0	0.0	98.1	.
0.020	1186	72768	8356	345	89.5	77.5	89.7	87.6	0.5
0.040	1033	77401	3723	498	94.9	67.5	95.4	78.3	0.6
0.060	963	78652	2472	568	96.3	62.9	97.0	72.0	0.7
0.080	912	79268	1856	619	97.0	59.6	97.7	67.1	0.8
0.100	851	79659	1465	680	97.4	55.6	98.2	63.3	0.8

Grupp II

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2451	0.1449	240.2318	<.0001
b3	1	0.3958	0.0742	28.4249	<.0001
b5	1	0.0760	0.0202	14.2267	0.0002
b10	1	0.1454	0.0125	136.0311	<.0001
e4	1	-0.00062	0.000242	6.5176	0.0107
e6	1	0.2836	0.0675	17.6725	<.0001
l6	1	0.9518	0.1486	41.0320	<.0001
l9	1	0.00406	0.00153	7.0338	0.0080
l10	1	0.00357	0.00125	8.2006	0.0042
l12	1	0.00281	0.00117	5.7954	0.0161
l16	1	0.000947	0.000554	2.9178	0.0876
l22	1	-0.00012	0.000041	8.6428	0.0033

l23	1	-0.00459	0.00252	3.3252	0.0682
l28	1	0.5970	0.1094	29.8087	<.0001
u3	1	0.000027	0.000013	4.4848	0.0342
u18	1	-0.00725	0.00136	28.2658	<.0001
RANTOMFR	1	1.5153	0.3672	17.0299	<.0001
ek1	1	-0.00010	0.000029	10.6684	0.0011
ek3	1	0.2022	0.0586	11.9193	0.0006
ek4	1	0.000083	0.000023	12.4454	0.0004
d4	1	0.3128	0.0823	14.4531	0.0001
a1	1	0.0511	0.0137	13.9891	0.0002

Association of Predicted Probabilities and Observed Responses

Percent Concordant	87.6	Somers' D	0.779
Percent Discordant	9.8	Gamma	0.799
Percent Tied	2.6	Tau-a	0.023
Pairs	19953172	c	0.889

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
18.1469	8	0.0202

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	556	0	35887	0	1.5	100.0	0.0	98.5	.
0.020	394	32180	3707	162	89.4	70.9	89.7	90.4	0.5
0.040	325	34536	1351	231	95.7	58.5	96.2	80.6	0.7
0.060	272	35052	835	284	96.9	48.9	97.7	75.4	0.8
0.080	252	35289	598	304	97.5	45.3	98.3	70.4	0.9
0.100	230	35423	464	326	97.8	41.4	98.7	66.9	0.9

Grupp III

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7311	0.1672	107.1461	<.0001
b3	1	0.2662	0.1520	3.0668	0.0799
b6	1	0.00403	0.000927	18.8909	<.0001
b10	1	0.0948	0.0170	31.1693	<.0001
e2	1	-0.0311	0.0162	3.6869	0.0548
e6	1	0.4776	0.1440	11.0077	0.0009
l5	1	0.7548	0.1723	19.1951	<.0001
l6	1	1.4770	0.1969	56.2564	<.0001
l9	1	0.0117	0.00254	21.0707	<.0001
l22	1	-0.00005	0.000034	2.4405	0.1182

l29	1	-0.00067	0.000182	13.6037	0.0002
l30	1	-0.0790	0.0470	2.8189	0.0932
u12	1	-0.00005	0.000018	8.4255	0.0037
u18	1	-0.0149	0.00243	37.6602	<.0001
u20	1	0.00374	0.00209	3.2062	0.0734
RANTOMFR	1	0.5436	0.2946	3.4061	0.0650
d4	1	0.4875	0.1514	10.3620	0.0013
a1	1	0.0394	0.0252	2.4418	0.1181

Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.8	Somers' D	0.711
Percent Discordant	12.8	Gamma	0.736
Percent Tied	3.4	Tau-a	0.017
Pairs	2886290	c	0.855

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
13.5402	8	0.0946

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	190	0	15191	0	1.2	100.0	0.0	98.8	.
0.020	109	13839	1352	81	90.7	57.4	91.1	92.5	0.6
0.040	75	14763	428	115	96.5	39.5	97.2	85.1	0.8
0.060	63	14944	247	127	97.6	33.2	98.4	79.7	0.8
0.080	57	15027	164	133	98.1	30.0	98.9	74.2	0.9
0.100	50	15066	125	140	98.3	26.3	99.2	71.4	0.9

Grupp IV

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5636	0.1773	77.7952	<.0001
b10	1	0.0412	0.0173	5.6646	0.0173
e1	1	-0.00073	0.000332	4.8337	0.0279
l1	1	-0.0751	0.0302	6.1665	0.0130
l6	1	1.8846	0.4239	19.7649	<.0001
l8	1	0.00884	0.00250	12.4552	0.0004
l29	1	-0.00043	0.000154	7.6406	0.0057
u18	1	-0.0214	0.00409	27.2575	<.0001
RANTOMFR	1	1.5434	0.3749	16.9435	<.0001
ek3	1	0.5300	0.1678	9.9729	0.0016
d4	1	0.8900	0.2125	17.5482	<.0001
a1	1	0.0381	0.0239	2.5394	0.1110

Association of Predicted Probabilities and Observed Responses

Percent Concordant	88.9	Somers' D	0.808
Percent Discordant	8.1	Gamma	0.833
Percent Tied	3.0	Tau-a	0.015
Pairs	617272	c	0.904

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
15.7533	8	0.0461

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	76	0	8122	0	0.9	100.0	0.0	99.1	.
0.020	49	7428	694	27	91.2	64.5	91.5	93.4	0.4
0.040	30	7883	239	46	96.5	39.5	97.1	88.8	0.6
0.060	26	7972	150	50	97.6	34.2	98.2	85.2	0.6
0.080	23	8015	107	53	98.0	30.3	98.7	82.3	0.7
0.100	21	8038	84	55	98.3	27.6	99.0	80.0	0.7

Grupp V

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3546	0.1320	105.2981	<.0001
b3	1	0.2777	0.1377	4.0700	0.0437
b5	1	0.1529	0.0513	8.8943	0.0029
b10	1	0.2043	0.0241	72.0107	<.0001
e2	1	-0.1379	0.0686	4.0402	0.0444
l1	1	-0.0961	0.0305	9.9312	0.0016
l5	1	0.8685	0.1059	67.2412	<.0001
l6	1	1.4326	0.0689	432.7390	<.0001
l8	1	0.00284	0.000852	11.0916	0.0009
l10	1	0.00498	0.000980	25.7956	<.0001
l13	1	0.00339	0.000786	18.5857	<.0001
l14	1	0.2929	0.1635	3.2086	0.0733
l18	1	0.00159	0.000380	17.5590	<.0001
l23	1	-0.00815	0.00265	9.4852	0.0021
l27	1	0.00487	0.00148	10.9142	0.0010
l29	1	-0.00205	0.000394	27.1549	<.0001
RANTOMFR	1	1.9649	0.4844	16.4524	<.0001
d1	1	0.3946	0.0809	23.8119	<.0001
d2	1	1.1319	0.1817	38.8050	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	89.8	Somers' D	0.805
Percent Discordant	9.3	Gamma	0.813
Percent Tied	0.9	Tau-a	0.090
Pairs	9308928	c	0.903

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
23.4037	8	0.0029

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	768	0	12121	0	6.0	100.0	0.0	94.0	.
0.020	706	6868	5253	62	58.8	91.9	56.7	88.2	0.9
0.040	618	10532	1589	150	86.5	80.5	86.9	72.0	1.4
0.060	573	11121	1000	195	90.7	74.6	91.7	63.6	1.7
0.080	551	11329	792	217	92.2	71.7	93.5	59.0	1.9
0.100	537	11448	673	231	93.0	69.9	94.4	55.6	2.0

Bilaga D

Validering (Skiljevärde på sannolikheterna inom parantes).

Probit, Grupp I (0.04)

The FREQ Procedure

Table of OBEST2 by pred_obest

OBEST2	pred_obest		
Frequency / Row Pct	0	1	Total
0	41204 95.57	1911 4.43	43115
1	259 34.17	499 65.83	758
Total	41463	2410	43873

Probit, Grupp II (0.02)

The FREQ Procedure

Table of OBEST2 by pred_obest

OBEST2	pred_obest		
Frequency / Row Pct	0	1	Total
0	17769 90.14	1944 9.86	19713
1	158 40.83	229 59.17	387
Total	17927	2173	20100

Probit, Grupp III (0.02)

The FREQ Procedure

Table of OBEST2 by pred_obest

OBEST2	pred_obest		
Frequency / Row Pct	0	1	Total
0	8531 92.50	692 7.50	9223
1	84 60.00	56 40.00	140
Total	8615	748	9363

Probit, Grupp IV (0,02)

The FREQ Procedure

Table of OBEST2 by pred_obest

OBEST2	pred_obest		
Frequency Row Pct	0	1	Total
0	4217 91.32	401 8.68	4618
1	24 70.59	10 29.41	34
Total	4241	411	4652

Probit, Grupp V (0,04)

The FREQ Procedure

Table of OBEST2 by pred_obest

OBEST2	pred_obest		
Frequency Row Pct	0	1	Total
0	5504 87.45	790 12.55	6294
1	106 26.63	292 73.37	398
Total	5610	1082	6692

Bilaga E

De viktigaste SAS-koderna använda i denna uppsats.

Exempel på Stegvis Logistisk regression:

```
proc logistic data=datamaterialet
descending/Utgår från utfall=1 istället för 0 i den beroende variabeln/
;
model Obest=variabler
/
lackfit           /hosmer and lemeshow´s goodness of fit test/
ctable             /tar fram klassificeringstabellen/
selection=stepwise /stegvis logistisk regression/
slentry=0.3        /signifikansnivå för variabel in i modellen/
slstay=0.3         /signifikansnivå för variabel ut ur modellen/
link=logit        /logit-regression/
;
output out=namn p=phat
/visar de beräknade sannolikheterna i vald fil tillsammans med den
ursprungliga datan/
run;
```

Test för multikollinearitet:

```
proc logistic descending data=sasbibl.ratz2;
model obest=variabler;
output out=sannolikheter pred=phat;
/logistisk regression för att få de skattade sannolikheterna/

data sannolikheter;
set sannolikheter;
w=phat*(1-phat);      /Skapar viktmatrisen/

proc reg data=sannolikheter;
weight w;
model obest=variabler
/
vif;                  /Variance Inflation Factor/
run;
```

Probit-regression:

```
proc logistic data=datamaterialet
descending
;
model Obest=variabler
/lackfit
ctable
link=probit        /probit-regression/
;
run;
```