Matematisk statistik
Stockholms universitet

Kreditvärdering med logistisk regression

Peter Zbornik

Examensarbete 2007:6

ISSN 0282-9169

Postadress:

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm
Sverige

Internet:

http://www.matematik.su.se/matstat

Matematisk statistik
Stockholms universitet
Examensarbete **2007:6**,
http://www.matematik.su.se/matstat

Kreditvärdering med logistisk regression

Peter Zbornik*

April 2007

Sammanfattning

Denna rapport beskriver utvecklingen av en logistisk regressionsmodell för kreditvärdering. Del I ger en historisk överblick av kreditvärdering. I Del II redovisas teorin bakom logistisk regression i samband med den generaliserade linjära modellen. Det motiveras att likelihood-ratiotestet och informationskriteria är ekvivalenta vid modellval. Accuracy Ratio används med korsvalidering som stoppkriterium för stepwisealgoritmen. Weight of Evidence används för att reducera antalet kategorier hos kategorialvariabler samt för varibeltransformation. Residualanalys diskuteras. Ovanstående idéer appliceras i Del III på ett dataset, där responsvariabeln mäter huruvida ett lån betalats tillbaka eller ej. Resultaten jämförs med stepwisealgoritmen i SAS. Stepwisealgoritmen med likelihood-ratiotest och korsvaliderat stoppkriterium ger goda resultat. Vinsten från den bästa modellen uppskattas.

*E-post: peter.zbornik@seznam.cz. Handledare: Thomas Höglund.

Abstract

This report describes the development of a credit scoring model using logistic regression. Part I gives a historic outline of credit scoring. Part II describes logistic regression in the context of the generalized linear model. The equivalence of the likelihood-ratio test to information criteria is motivated. Accuracy Ratio is used with cross-validation as a stop-criterion for the stepwise algorithm. Weight of Evidence is used to reduce the number of categories in categorical variables and for variable transformation. The analysis of residuals is discussed. The above-mentioned ideas are in Part III applied to a dataset, where the response variable indicates whether a loan was repaid or not. The results are compared with the stepwise-algorithm in SAS. The stepwise algorithm with likelihood-ratio tests and cross-validated stop-criterion shows good performance. Profitability from the best model is estimated.

Förord

Denna rapport utgör mitt examensarbete i Matematisk Statistik på 20 poäng från Stockholms Universitet, och avslutar min magisterutbildning på Matematisk-Datalogisk linje.

Stort tack till Helena för stort tålamod och till Thomas Höglund på Stockholms Universitet för insiktsfull handledning.

Innehåll

Inledning

Ordet ”kredit” kommer från latinets ”credo”, som betyder ”att tro”. En långivare lånar ut pengar till en låntagare i tro att han kommer betala tillbaka lånet. Låntagaren betalar emellertid inte alltid tillbaka, så långivaren försöker därför uppskatta risken av att låntagaren inte betalar tillbaka lånet innan pengarna är utlånade. Långivaren beviljar endast lånet till dem, som han tror kommer betala tillbaka tillräckligt mycket av lånet för att han ska gå med vinst.

Statistiska modeller används flitigt inom finanssektorn. Ett av tillämpningsområdena är credit scoring, eller kreditvärdering. Kreditvärdering syftar till att förutsäga sannolikheten att en klient kommer att betala tillbaka det lån han söker till. Förbättrad kreditvärdering ger företaget högre inkomster för samma låneprodukt, eftersom företaget antingen kan acceptera fler ansökningar med bibehållen risknivå eller sänka förlusterna med bibehållen acceptationsnivå. Företag med god kreditvärdering kan tillåta sig att erbjuda en enskild konsument lån med lägre ränta än konkurrenterna med samma vinstmarginal som konkurrenterna. På sikt kan förbättrad kreditvärdering förväntas ge lägre ränta för konsumenten.

Del I. Kreditvärderingens historia

Kreditmarknaden fram till 1950

I början 1800-talet var största delen av Sveriges invånare bönder. Banker beviljade endast lån till välbärgade. Institutionella lån från olika kassor, som fattigkassan eller kyrkokassan, förekom i liten skala.

1800-talets lantbrukare lånade mest inom socknen eller av den lokala lanthandlaren. Lantbrukarna brukade inte flytta och handlaren hade det svårt att vinna nya kunder utanför sitt handelsområde, då vägar och kommunikationer fungerade dåligt. När lantbrukarna behövde pengar till utsäde på våren gav lanthandlaren eller en vän kredit. Efter skörden på hösten betalade man tillbaka, förutsatt att skörden blev den förväntade. Detta samarbete fungerade tack vare insikten om att man hade ett gemensamt hjälpbehov. Det var vanligt att man både var gäldenär och borgenär på samma gång. Indrivning skedde sällan.

Handlarna och köpmännen beviljade nästan uteslutande kredit till folk de kände väl. De hade förstahandsinformation om sina kunders finanser, och kunde vara säkra på vem som hade god riskprofil. När skörden slog fel, eller gäldenären dog, hände det dock ofta att handlaren gick i konkurs.

Från slutet av 1800-talet till 1950, skedde en massiv inflyttning till storstäderna. Andelen jordbrukare minskade och andelen arbetare i tillverkningsindustrin växte kraftigt. Sverige gick från att vara ett fattigt lantbruksland till att bli ett välmående folkhem.

1800-talets kreditmarknad var baserad på ömsesidigt beroende, god insikt i låntagarens finanser, säsongsbundna inkomster och på en geografiskt begränsad marknad. I och med den stora inflyttningen till städerna och den ökade mobiliteten bröts som industrialiseringen innebar slutade lånemodellen från 1800-talet att fungera. Långivarens insyn i låntagarens ekonomi var begränsad och inkomsterna var inte längre säsongsbundna.

Under början av 1900-talet var de vanligaste kreditinstitutionerna pantbanken och den lokala lånehajen. Pantbanken gav lån mot säkerhet. Lantbrukarens höstskörd ersattes med silver, smycken eller andra ägodelar. Lånehajen lånade ut pengar utan säkerhet men till hög ränta. Den som inte betalade kunde råka illa ut.

Allt eftersom levnadsstandarden ökade i Sverige, stadsbefolkningen stabiliserades och masstillverkningen tog fart, uppkom också en reguljär lånemarknad. Efter kriget blev det vanligare att ta amorteringslån för möbel- och fordonsinköp. Räntenivån för lån utan säkerhet var fortfarande hög.

Statistisk kreditvärdering

Grundaren till statistiskt baserad kreditvärdering i USA anses vara David Durand. I sin artikel från 1941, Risk elements in consumer instalment financing (Durand, 1941) använde Durand diskriminantanalys för att bedöma risk för betalningsförsummelse hos konsumentlån.

Diskriminantanalysen hade utvecklats av Ronald Fisher[1] fem år tidigare när han studerade karakteristiska parametrar för olika typer av iris hos blommor och den geografiska härkomsten för olika typer av kranium.

Vid samma tid hade postorderfirmor och finansföretag problem med att bedöma låneansökningar. Flertalet riskanalytiker hade lämnat sina arbeten för att tjäna som soldater under andra världskriget. Det fanns inte tillräckligt med folk som kunde bedöma betalningsförmågan hos klienterna. Innan riskanalytikerna gick ut i kriget skrev de ned tumregler för kreditbedömning. Dessa regler användes sedan av lekmän för att bevilja eller avvisa kreditansökningar, och utgjorde ett av de första exemplen på ett expertsystem.

Ronald Fisher 1890-1962

1956 grundades det första konsultbolaget för riskvärdering av Bill Fair och Earl Isaac, som sålde tumregler för kreditvärdering baserade på diskriminantanalys till postorder- och finansföretag, så kallade scorekort.

Förklarande parametrar i dessa scorekort var t.ex. inkomst, ålder och utbildning. Varje nivå hos respektive parameter hade ett antal poäng, som representerade vikterna i diskriminantfunktionen. Poängen sammanräknades, och om poängsumman var högre än en bestämd konstant, en s.k. cut-off, var risken för utebliven återbetalning acceptabel och lånet beviljades.

Tabell 1 är ett exempel på ett scorekort. Enligt detta scorekort får en 30-årig manlig akademiker 50 poäng (20 för ålder, 30 för yrke och 0 för kön), vilket är mer än cut-off på 40 poäng. Hans låneansökan blir sålunda beviljad.

En 20-årig kvinnlig försäljare får enligt samma tabell 25 poäng. Hennes ansökan blir avvisad.

Tabell 1. Exempel på ett scorekort

Parameter	Intervall	Poäng
Ålder	18-25	0
	26-40	20
	41-55	30
	56-65	25
Yrke	Arbetare	0
	Akademiker	30
	Tjänstesektorn	25
	Försäljning	15
Kön	Man	0
	Kvinna	10
	CUTOFF	40

Statistiska scorekort ersatte riskbedömningen som långivaren tidigare var utförde med hjälp av sitt goda omdöme. Riskanalyser massproducerades och antalet försummade lån sjönk med 50% till följd av den förbättrade kvaliteten på riskanalysen, som scorekorten medförde. Räntenivån sänktes med p.g.a. den lägre risknivån.

Kreditbyråer var den andra nödvändiga förutsättningen för att man i den industrialiserade världen skulle kunna erhålla samma information som handelsmannen i byn tidigare hade tillgänglig för att bedöma om han skulle låna ut pengar till en kund eller ej. Det fanns gott om kunder som inte betalade tillbaka sina lån. Dessa kunde ta lån i olika banker, utan att betala tillbaka. Tidigt uppkom en insikt om att kreditinstitutionerna hade ett gemensamt intresse av att dela med sig information om dåliga och bra kunder.

De första större kreditregistren uppkom i USA på 1960-talet. De utgjorde en lösning som kopierade den lokala handelsmannens information om byinnevånarna. Kreditregistren var den moderna världens svar på byskvallret. Personer som inte betalade tillbaka sitt lån registrerades

[1] Bilden på Ronald Fisher är hämtad från: http://en.wikipedia.org/wiki/Image:RonaldFisher.jpg

i kreditregistret. När samma person sökte sitt nästa lån kunde låneinstitutionen begära ut en registerutskrift som visade att kunden hade försummat att betala sitt tidigare lån. En sådan person fick det svårt att få sitt nästa lån beviljat. Å andra sidan var det lättare att få lån om man betalat sina tidigare lån utan dröjsmål.

På 80-talet snabbades kreditbedömningsprocessen upp med datorernas intrång. Scorekort automatiserades och kunden kunde omedelbart få besked om huruvida hans lån var beviljat eller ej. Diskriminantanalysen byttes ut mot den något robustare men mer beräkningsintensiva logistiska regressionsmodellen, vilken utgör fokus för denna rapport.

Det tog det industrialiserade samhället ett sekel att ersätta bygemenskapens och den lokala köpmannens roll i det agrara samhället.

Del II. Teoretiska resultat

I denna del visar vi teoretiska resultat, för den logistiska regressionsmodellen. Logistisk regression används för att förutsäga en händelse, t.ex. att klienten inte betalar tillbaka sitt lån i tid.

Vi går igenom teorin den logistiska regressionsmodellen som ett specialfall av den generaliserade linjära modellen. Vidare diskuteras verktyg, som används tillsammans med den logistisk regression: modellval, korsvalidering, residualer samt förbehandling av kategorialvariabler. Särskilt inriktar vi oss på modifieringar av stepwisealgoritmen i SAS. Verktygen från Del II används sedan i Del III för att utveckla en kreditvärderingsmodell på ett dataset med information om personlån.

Logistisk regression

Logistisk regression använder sig i likhet med linjär regression av träningsdata, $\mathbf{X}$, som beskriver omständigheterna för ett antal händelser $\mathbf{Y}$. $\mathbf{X}$ och $\mathbf{Y}$ innehåller observationer av ett antal variabler. Variablerna i $\mathbf{X}$ kallas för förklarande variabler eller parametrar och variabeln i $\mathbf{Y}$ kallas för responsvariabel. Utifrån $\mathbf{X}$ och $\mathbf{Y}$ försöker vi konstruera en modell, som gör en kvalificerad gissning av sannolikheten för att en ny händelse inträffar, givet specifika värden på de förklarande variablerna.

De förklarande parametrarna i $\mathbf{X}$ kan exempelvis innehålla parametrar rörande en låneansökan, som yrke, ålder och inkomst, och responsvariabeln i $\mathbf{Y}$ kan anta värdet 1 om lånet avslutats p.g.a. betalningsförsummelse och 0 annars. Vi söker att så bra som möjligt förutsäga sannolikheten för att responsvariabeln antar värdet 1, givet några värden de förklarande variablerna.

Formellt har vi en $(n \times k)$ datamatris $\mathbf{X}$, och en $(n \times 1)$ kolumnvektor $\mathbf{Y}$. $\mathbf{X}$ innehåller k parametrar och n oberoende observationer av dessa parametrar. $\mathbf{Y}$ innehåller n oberoende observationer av utfallet från en stokastisk variabel. Mot observation nummer i svarar en rad i $\mathbf{X}$, kallad $\mathbf{X}_i$ och en observation y_i i $\mathbf{Y}$.

Vi kommer att koncentrera oss på att finna sannolikheten p_i för en händelse $Y_i = y_i$, $1 \le i \le n$, $y_i \in \{0,1\}$, givet en situation, som beskrivs av parametrarna i radvektorn $\mathbf{X}_i$.

Vi söker med andra ord finna:

$$p_i \equiv \mathrm{E}\left(Y_i = y_i \middle| \mathbf{X}_i\right), \text{ där}$$

$$Y_i \sim p_i^{y_i}(1-p_i)^{(1-y_i)} \equiv Be(p_i), \text{ och } Y_i \text{ är oberoende, givet } \mathbf{X}_i. \quad (1).$$

Från Bernoullifördelningen har vi att $\mathrm{Var}\left(Y_i\right) = p_i(1-p_i)$.

En naiv ansats för att försöka förutsäga p_i är en linjär regressionsmodell:

$$p_i = \mathbf{X}_i\beta, \quad (2)$$

där $\beta \equiv (\beta_1,...,\beta_k)^T$ är en kolumnvektor med *k* parametrar.

Det gäller således att (2) kan skrivas som $p_i \equiv \beta_1\mathbf{X}_{i,1} + \beta_2\mathbf{X}_{i,2} + ... + \beta_k\mathbf{X}_{i,k}$, där $\mathbf{X}_{i,j}$ anger värdet för rad *i*, kolumn *j* i $\mathbf{X}$.

Ansats (2) har dock den kontraintuitiva egenskapen, att för stora, eller små värden på $\mathbf{X}_i$, kan det hända att skattningen $\hat{p}_i$ är större än ett eller mindre än noll, vilket borde vara omöjligt.

Vi försöker istället finna en länkfunktion *g,* sådan att det aldrig kan inträffa att $p_i < 0$ eller $p_i > 1$, oavsett hur stora eller små värdena i $\mathbf{X}_i$ är. Vi ändrar således (2) till:

$$g(p_i) = \mathbf{X}_i\beta$$

Tre funktioner används ofta som länkfunktion *g*[2]:

1. Logitfunktionen, $\text{logit}(p_i) \equiv \ln\left(\frac{p_i}{1-p_i}\right)$, (3)
2. Komplementära log-logfunktionen, $\ln(-\ln(1-p_i))$
3. Inversen till fördelningsfunktionen (kvantilfunktionen) för normalfördelningen med $\mu_i = 0$ och $\sigma^2 = 1$: $\text{probit}(p_i) \equiv \Phi^{-1}(p_i)$,

För alla dessa funktioner gäller att dess invers g^{-1} har värdemängden $(0,1)$. Då $p_i = g^{-1}(\mathbf{X}_i\beta)$, kan situationen att $p_i < 0$ eller $p_i > 1$ aldrig infinna sig. Vi väljer logitfunktionen i (3) ovan som länkfunktion.

Logitfunktionen (3) har följande önskvärda egenskaper:

- Oddsfunktionen kan tolkas som $\frac{\text{P(händelse)}}{\text{P(ingen händelse)}}$
- Logitfunktionen är symmetrisk: modellen för P(händelse) är ekvivalent med modellen för P(ingen händelse) med omvänt tecken på β
- Koefficienterna β betecknar hur log-oddset förändras när parametrarna i X ökar med ett.
- Summan av alla residualer för en modell med intercept är noll.[3]

[2] Valet av länkfunktion kan göras mer komplicerat, se Fahrmeir och Tutz (2001) för en introduktion.
[3] Gäller för alla generaliserade linjära modeller med intercept och kanonisk länkfunktion, se Woods (2006).

Vi är intresserade av att ha ett intercept i modellen, så vi låter alltid fortsättningsvis kolumn 1 i $\mathbf{X}$, bara innehålla 1:or, d.v.s. $p_i = \beta_1 + \beta_2 \mathbf{X}_{i,2} + ... + \beta_k \mathbf{X}_{i,k}$.

Den logistiska regressionsmodellen vi kommer arbeta i fortsättningen med har således följande form:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{X}_i\beta \ , p_i \equiv \mathrm{E}(Y_i),\ Y_i \sim Be(p_i),\ Y_i \text{ oberoende givet } \mathbf{X}_i,\ \beta \equiv \{\beta_1,...,\beta_k\}^T \quad (4)$$

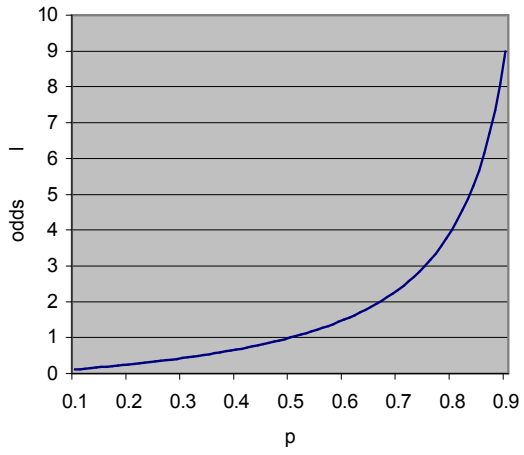
Det följer att

$$p_i = g^{-1}(\mathbf{X}_i\beta) = \frac{\exp(\mathbf{X}_i\beta)}{1+\exp(\mathbf{X}_i\beta)} = \frac{1}{1+\exp(-\mathbf{X}_i\beta)}. \quad (5)$$

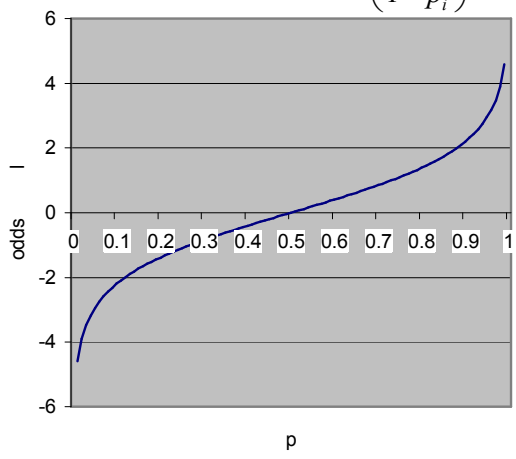
Efter att ha skattat β kan vi från (5) skatta p_i givet $\mathbf{X}_i$.

Oddsfunktionen, logitfunktionen och den inversa logitfunktionen visas i Figur 1a, 1b och 1c nedan.

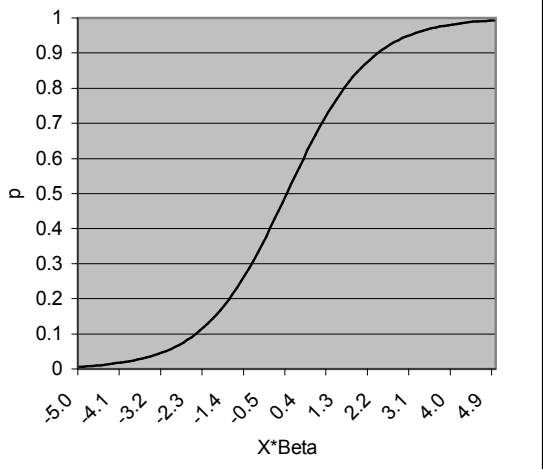
Figur 1a. Oddsfunktionen: $\frac{p_i}{1-p_i}$.



Figur 1b. Logitfunktionen: $\ln\left(\frac{p_i}{1-p_i}\right)$.



Figur 1c. Inversa logitfunktionen:
$p_i = g^{-1}(\mathbf{X}_i\beta)$



Värt att nämnas är att variansen i den linjära regressionsmodellen är konstant för alla observationer, medan variansen för den logistiska regressionsmodellen (1), $p_i(1-p_i)$ minskar med avståndet från p=0,5.

Den generaliserade linjära regressionsmodellen

Den logistiska regressionsmodellen är ett specialfall av den generaliserade linjära regressionsmodellen:

$$g(\mu_i) = \mathbf{X}_i\beta\,,\ 1 \le i \le n\,, \quad (6)$$

där

- $\mu_i \equiv \mathrm{E}(Y_i)$,
- Y_i är oberoende stokastiska variabler, givet $\mathbf{X}_i$.
- $Y_i \sim \exp\left[\left\{y_i\theta_i - \mathrm{b}(\theta_i)\right\}/\mathrm{a}(\phi) + \mathrm{c}(y_i,\phi)\right]$, vilket betecknar den kanoniska formen för den exponentiella familjen. Denna familj innehåller många kända fördelningar som bland annat normal-, binomial-, poisson- och gamma-fördelningarna. θ_i kallas för den kanoniska parametern.
- β är en kolumnvektor $\{\beta_1,...,\beta_k\}^T$ med *k* okända parametrar,
- *g* är en monoton kontinuerligt differentierbar funktion. *g* kallas för länkfunktion.

För den generaliserade linjära modellen gäller följande resultat för väntevärde och varians:

$$\mathrm{E}(Y_i) \equiv \mu_i = \mathrm{b}'(\theta_i)\,,\ \mathrm{Var}(Y_i) = \mathrm{b}''(\theta_i)\,\mathrm{a}(\phi)\,. \quad (7)$$

Vi motiverar nedan att den logistiska regressionsmodellen (4) uppfyller villkoren för generaliserad linjär modell enligt (6) ovan.

Först visas att Bernoullifördelningen i (1) och (4) är tillhör den exponentiella familjen:

$$p^{y_i}(1-p_i)^{(1-y_i)} = \exp\left[\ln(p_i^{y_i}(1-p_i)^{(1-y_i)}\right] = \exp\left[y_i \ln(p_i) + (1-y_i)\ln(1-p_i)\right]$$

$$= \exp\left[y_i \ln\left(\frac{p_i}{1-p_i}\right) + \ln(1-p_i)\right] = \exp\left[\left\{y_i\theta_i - \mathrm{b}(\theta_i)\right\}/\mathrm{a}(\phi_i) + \mathrm{c}(y_i,\phi_i)\right],$$

$$\theta_i \equiv \ln\left(\frac{p_i}{1-p_i}\right) \equiv \mathrm{logit}\left(p_i\right),\ \mathrm{b}(\theta_i) \equiv \ln(1+e^{\theta_i}),\ \mathrm{a}(\phi_i)=1,\ \mathrm{c}(y_i,\phi_i) \equiv 0. \quad (8)$$

Logitfunktionen (3) uppfyller villkoret för monotonicitet och deriverbarhet i (6). Monotonicitet följer från att oddsfunktionen och logaritmfunktionen båda är monotont växande. Kontinuerlig deriverbarhet får vi från faktumet att derivatan av $\ln(x)$ är x^{-1} och att oddsfunktionen aldrig antar värdet 0.

Från (7) har vi de kända resultaten att

$$\mathrm{b}'(\theta_i) = \mu_i = p_i \text{ och } \mathrm{b}''(\theta_i) = \mathrm{Var}(Y_i) = p_i(1-p_i). \quad (9)$$

En annan iakttagelse är att den kanoniska parametern θ_i i (8) är ekvivalent med länkfunktionen *g*. I dessa fall kallas länkfunktionen *g* för kanonisk länkfunktion. Logitfunktionen (3) är således den kanoniska länkfunktionen till Bernoullifördelningen.

Maximum-likelihoodmetoden

I vår modell (4) är parametrarna β inte kända. Dessa parametrar måste således skattas, för att vi ska kunna beräkna skattningen av sannolikheten $\hat{p}_i \equiv \mathrm{P}(Y_i=1)$ från inversen till den kanoniska länkfunktionen (5).

Vi kallar skattningen till β för $\hat{\beta}$. Nedan visas några olika skattningsmetoder och ansatser för att finna $\hat{\beta}$:

- Maximum-likelihoodmetoden,
- Minsta kvadratmetoden,
- Bayes-skattningar,
- Momentmetoden,
- MCMC-skattningar,
- Bootstrapping,
- Rao-Blackwell skattningar.

Vi kommer att använda oss av maximum-likelihoodmetoden som skattningsmetod för $\hat{\beta}$.

Vi definierar likelihoodfunktionen *L,* som följande funktion:

$$L \equiv L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}|\boldsymbol{y}) \equiv f_{\boldsymbol{\theta}}(\boldsymbol{y}), \quad (10)$$

där $1 \le i \le n,\ L : \Theta \mapsto \mathbb{R},\ \boldsymbol{\theta} = \{\theta_1,...,\theta_k\}^T \in \Theta,\ \boldsymbol{y} = \{y_1,...,y_n\}^T$, $\mathbf{Y} \sim f_{\boldsymbol{\theta}}(\boldsymbol{y})$. Vi kallar funktionen $L(\boldsymbol{\theta}|\boldsymbol{y})$ för likelihoodfunktion, eftersom den är definierad på parameterrummet Θ, till skillnad från fördelningsfunktionen $f_{\boldsymbol{\theta}}(\boldsymbol{y})$ som är definierad på mängden av alla utfall $\boldsymbol{y}$ från den stokastiska variabelvektorn $\mathbf{Y}$. $\boldsymbol{\theta}$ är en parameter och inte en stokastisk variabel. Vi kommer i fortsättningen omväxlande att beteckna likelihoodfunktionen som $L,\ L(\boldsymbol{\theta})$ eller $L(\boldsymbol{\theta}|\boldsymbol{y})$.

Maximum-likelihoodmetoden söker finna den skattning $\hat{\boldsymbol{\theta}}$, som maximerar likelihoodfunktionen:

$$L(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) \ge L\left(\hat{\tilde{\boldsymbol{\theta}}}|\boldsymbol{y}\right),\ \hat{\tilde{\boldsymbol{\theta}}} \in \Theta$$

Det kan visas att maximum-likelihoodskattningen (ML-skattningen) är *invariant:*

$$\text{om } \hat{\boldsymbol{\theta}} \text{ är en ML-skattning, så är } \hat{\gamma} \equiv g(\hat{\theta}) \text{ också en ML-skattning.} \quad (11)$$

Egenskapen (11) ger oss genom inversen till länkfunktionen i (5) att $\hat{p}_i$ är en ML-skattning, om $\hat{\beta}$ är en ML-skattning.

Vidare kan man för ML-skattningar under milda villkor för den generaliserade linjära modellen visa[4], att om: (12)

1. log-likelihood funktionen, ln(L) är två gånger deriverbar och
2. $I \ne 0,\ I(\boldsymbol{\theta}) \equiv -\mathrm{E}\left[\dfrac{\partial^2 \ln(L)}{\partial \boldsymbol{\theta}^2}\right]$. $I(\boldsymbol{\theta})$ kallas för Fishers informationsmatris,

så gäller att $\hat{\boldsymbol{\theta}}$ är en asymptotiskt optimal skattning i följande bemärkelse:

1. $\hat{\boldsymbol{\theta}}$ är asymptotiskt *konsistent*: Om $\hat{\boldsymbol{\theta}}_n$ är baserad på n observationer $y_1,...,y_n$ och $\boldsymbol{\theta}_0$ är det sanna värdet av $\boldsymbol{\theta}$,
   så gäller att $\lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right| < \varepsilon\right) \to 1,$ för alla positiva ε, (13)
2. $\hat{\boldsymbol{\theta}}$ är asymptotiskt $\sim N\left(\boldsymbol{\theta}_0, I(\boldsymbol{\theta})^{-1}\right)$, där $I(\boldsymbol{\theta})$ är Fishers informationsmatris. (14)
3. $\hat{\boldsymbol{\theta}}$ är asymptotiskt *effektiv*: ingen estimator har asymptotiskt lägre varians än maximum-likelihoodskattningen $\hat{\boldsymbol{\theta}}$. (15)

Ovanstående egenskaper gör Maximum-likelihoodmetoden till en lämplig skattningsmetod för den logistiska regressionsmodellen.

[4] Villkoren gäller för alla normala modeller, förutsatt att antalet förklarande variabler i $\mathbf{X}$ inte växer för fort med antalet observationer n. Se Fahrmeir och Tutz (2001) för en utförligare diskussion.

$\hat{\boldsymbol{\theta}}$ kan vanligtvis inte uttryckas analytiskt, utan får uppskattas med hjälp av numeriska metoder.

Maximum-likelihoodmetoden för logistisk regression

Maximum-likelihoodmetoden för att skatta β i (4) kan delas upp i två steg:

1. Definition av likelihoodfunktionen
2. Beräkning av det $\hat{\beta}$ som maximerar likelihoodfunktionen

Vi avslutar kapitlet med att beskriva de asymptotiska resultaten (13), (14) och (15) för den logistiska regressionsmodellen (4).

Definition av likelihoodfunktionen för den logistiska regressionsmodellen

Låt oss börja med att definiera likelihoodfunktionen för β i den logistiska regressionsmodellen (4). Vi har från (4) att $Y_i \sim Be(p_i)$, Y_i oberoende. Vi söker således maximum för:

$$L \equiv L(\beta) \equiv L(\beta|\boldsymbol{y}) = f_\beta(\boldsymbol{y}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} = \prod_{i=1}^{n}\left(\frac{p_i}{1-p_i}\right)^{y_i}(1-p_i), \quad (16)$$

där den första likheten följer från (10), och den andra likheten följer från att Y_i är oberoende. Vi utnyttjar att $\max(L) = \max(\ln(L))$ och väljer att maximera

$$\ln\left(L(\beta|\boldsymbol{y})\right) = \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}\ln(1-p_i) = \sum_{i=1}^{n} y_i \mathbf{X}_i\beta - \sum_{i=1}^{n}\ln\left(1+e^{\mathbf{X}_i\beta}\right), \quad (17)$$

där den sista likheten följer från logitfunktionen i (4).

Maximering av likelihoodfunktionen för den logistiska regressionsmodellen

Olika metoder kan användas för att finna maximum av en funktion. En vanlig metod är att sätta de partiella derivatorna för varje parameter i β till noll, och sedan lösa ut värdet för β iterativt. Komponent *j* i vektorn av de partiella derivatorna till ln(*L*) är:

$$\frac{\partial \ln(L)}{\partial \beta_j} = \sum_{i=1}^{n} y_i \mathbf{X}_{i,j} - \sum_{i=1}^{n} \mathbf{X}_{i,j}\frac{1}{\left(1+e^{-\mathbf{X}_i\beta}\right)} = \sum_{i=1}^{n}\mathbf{X}_{i,j}(y_i - p_i) \equiv J_j(\beta),\ 1 \le j \le k.$$

där $\mathbf{X}_{i,j}$ betecknar värdet för rad *i*, kolumn *j* i modellmatrisen $\mathbf{X}$. Den första likheten är derivatan av (17). Den andra likheten följer från (5). Vi får att:

$$\frac{\partial \ln(L)}{\partial \beta} = \mathbf{X}^T(\mathbf{Y} - \mathbf{P}) \equiv J(\beta),\ \mathbf{P} \equiv (p_1, ..., p_n)^T$$

Då β är en vektor med *k* parametrar har vi således *k* ekvationer att lösa.

β varierar icke-linjärt med $\mathbf{X}$ och $\mathbf{Y}$, och går ej att lösa ut analytiskt, till skillnad från normalekvationerna i linjär regression. Vi använder oss istället av numeriska metoder för att finna maximum av likelihoodfunktionen (16).

En vanlig metod för att finna maximum av likelihoodfunktionen kallas för Fisher scoring. Fisher scoring är identisk med Newton-Raphson algoritmen för binär respons och logitfunktionen som länkfunktion, dvs den typ av logistisk regression som vi behandlar i denna rapport. Vi beskriver Newton-Raphson algoritmen nedan. Alogritmen uppdaterar skattningen $\hat{\beta}$ iterativt enligt nedan:

$$\hat{\beta}_{j+1} = \hat{\beta}_j - I^{-1}(\hat{\beta}_j) J(\hat{\beta}_j),$$

där *I* är Fishers informationsmatris från (12).

Som initialskattning i algoritmen sätter vi $\hat{\beta}_0 = \mathbf{0}$. I SAS anses algoritmen konvergerat, när

$$\frac{J(\hat{\beta}_j)^T I^{-1}(\hat{\beta}_j) J(\hat{\beta}_j)}{\left|\ln\left(L(\hat{\beta}_j)\right)\right| + 10^{-6}} < 10^{-8}$$

Den konvergerade skattningen är vår ML-skattning av β. Låt oss beteckna den med $\hat{\beta}_{\mathrm{ML}}$.

Asymptotiska resultat för den logistiska regressionsmodellen

Man kan visa att villkoren i (12) är uppfyllda för den logistiska regressionsmodellen.[5] Nedan skisseras beviset. Vi börjar med att visa att ln(*L*) i (17) är två gånger deriverbar:

$$\frac{\partial^2 \ln(L)}{\partial \beta^2} = \frac{\partial J(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta} \mathbf{X}^T \left(\mathbf{Y} - \left(1 + e^{-\mathbf{X}\beta}\right)^{-1}\right) = -\mathbf{X}^T \frac{\partial}{\partial \beta}\left(1 + e^{-\mathbf{X}\beta}\right)^{-1} =$$

$$-\mathbf{X}^T \left(-\left(1 + e^{-\mathbf{X}\beta}\right)^{-2}\right)\left(-\mathbf{X}^T e^{-\mathbf{X}\beta}\right) = -\mathbf{X}^T \mathbf{P}(1 - \mathbf{P})\mathbf{X} = -\mathbf{X}^T \operatorname{Var}(\boldsymbol{Y})\mathbf{X}$$

$$\Rightarrow I(\beta) = \mathbf{X}^T \operatorname{Var}(\boldsymbol{Y})\mathbf{X} \equiv \mathbf{X}^T \mathbf{V} \mathbf{X} \quad (18)$$

[5] Se Fahrmeir och Tutz (2001)

där $\boldsymbol{Y} \equiv (Y_1, ..., Y_n)^T$. Då Y_i i $\boldsymbol{Y}$ är oberoende gäller att $\mathrm{Var}(\boldsymbol{Y}) \equiv \mathbf{V}$ är en diagonalmatris med $\mathbf{V}_{i,i} = \mathrm{Var}(Y_i) = p_i(1-p_i)$. Log-likelihood funktionen är således två gånger deriverbar. $I(\beta) \neq 0$ om inte p_i=0 eller p_i=1 för alla *i*.

Resultaten (13), (14) och (15) är således tillämpbara.

Vi sammanfattar med att $\hat{\beta}_{\mathrm{ML}}$ för den logistiska regressionsmodellen är asymptotiskt väntevärdesriktig, effektiv och normalfördelad med

$$\hat{\beta}_{\mathrm{ML}} \sim N\left(\beta, \left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\right) \quad (19)$$

De asymptotiska resultaten ovan innebär att vi för ”stora” *n* kan skatta konfidensintervall för β och följdaktligen även för $p_i = \dfrac{1}{1+\exp(-\mathbf{X}_i\beta)}$. Dessa konfidensintervall är dock inte effektiva om vi inte vårt *n* är stort. Test baserade på normalfördelningen i (19) kallas för Wald-test och kan även användas för signifikanstest och konfidensintervall av enskilda parameterskattningar i $\hat{\beta}$.

Likelihood-ratiotestet

Ofta vill vi testa om en modell blir bättre i och med att vi lägger till en parameter. Till det kan vi använda oss av konfidensintervallen i Wald-testet ovan. I litteraturen föredras dock ett annat test framför Wald-testet (se Hosmer och Lemeshow (2000)). Detta test kallas för Likelihood-ratiotestet och beskrivs nedan.

Antag att vi vill testa hypotesen

$$\mathrm{H}_0 \ : \ \mathrm{g}(\boldsymbol{\mu}) = \mathbf{X}_0\beta_0$$

mot hypotesen

$$\mathrm{H}_1 \ : \ \mathrm{g}(\boldsymbol{\mu}) = \mathbf{X}_1\beta_1,$$

där H_0 och H_1 beskriver två generaliserade linjära modeller (6) , $\boldsymbol{\mu}$ är väntevärdesvektorn för en responsvektor $\mathbf{Y}$, och där $\mathbf{X}_0$ och $\mathbf{X}_1$ är två $(n \times m_0)$ resp $(n \times m_1)$ modellmatriser för vilka gäller att $\mathbf{X}_0 \subset \mathbf{X}_1$.

Låt vidare $\ln\left(L\left(\hat{\beta}_0\right)\right)$ och $\ln\left(L\left(\hat{\beta}_1\right)\right)$ vara de ML-maximerade log-likelihood funktionerna från föregående kapitel. Om H_0 är sann, så gäller att:

$$-2\lambda_{0,1} \equiv -2\left[\ln\left(L\left(\hat{\beta}_0\right)\right) - \ln\left(L\left(\hat{\beta}_1\right)\right)\right] \sim \chi^2_{m_1-m_0} \ . \quad (20)$$

Vi kan således testa om någon parameter β_a i β_1 är noll genom att definiera β_0 som parametervektorn β_1 utan denna parameter.

Det är värt att notera att (20) kan skrivas som ett uttryck innehållande kvoten mellan $L\left(\hat{\beta}_1\right)$ och $L\left(\hat{\beta}_0\right)$:

$$-2\lambda_{0,1} \equiv -2\ln(\Lambda),\ \Lambda \equiv \frac{L\left(\hat{\beta}_0\right)}{L\left(\hat{\beta}_1\right)},$$

därav namnet likelihood-ratiotestet. $-2\lambda_{0,1}$ är således stor, när kvoten Λ är liten.

I den logistiska regressionsmodellen använder vi log-likelihooden från (17) till detta test.

Modellval

Modellval utförs för att undvika att brusvariabler kommer in i modellen och försvagar dess prediktiva förmåga.

Frågan är vilka variabler som ska väljas till modellen. Antalet möjliga variabelkombinationer växer fort. För ett dataset med 50 variabler är antalet möjliga modeller $= \sum_{i=1}^{50} \binom{50}{i} = 2^{50} - 1 \approx 10^{15}$. Det är omöjligt att undersöka alla dessa modeller, så ett alternativt tillvägagångssätt krävs. Vi kommer att använda oss av en metod för modellval, som kallas för stepwise selection eller stepwisealgoritmen. Denna metod kan också lätt ta med brusvariabler, vilket gäller för alla metoder för modellval som finns tillgängliga i SAS®. Det gäller även att p-värdena för variablerna inte är exakta, utan endast anger den relativa vikten hos en variabel jämfört med de andra variablerna. I Del III förbehandlar vi de förklarande variablerna och använder oss av stepwisealgoritmen med likelihood-ratiotestet och korsvalidering för att minska risken för att brusvariabler kommer med i den slutliga modellen.

Stepwisealgoritmen [6]

Detta avsnitt skisserar stepwisealgoritmen som en metod för modellval.

Stepwisealgoritmen:

- Antag att vi har en generaliserad regressionsmodell med en modellmatris **X**, där vi lagt till en parameter som antar värdet 1 för alla observationer.
- Låt parametervektorn till **X** vara β.
- Definiera $\hat{\beta}_r$, som en parametervekor vid iterationssteg *r*, med enskilda parameterskattingar $\hat{\beta}^{(j)}$, där *j* anger att $\hat{\beta}^{(j)}$ inkluderades i $\hat{\beta}_r$ vid iterationssteget *j*, $1 \le j \le r$. Antalet parametrar i $\hat{\beta}_r$ är mindre än eller lika med *r*. Exempelvis kan

[6] Se Hosmer och Lemeshow (2000) för en utförligare diskussion

vi ha att $\hat{\beta}_4 \equiv \left(\hat{\beta}^{(1)}, \hat{\beta}^{(4)}, 0, \hat{\beta}^{(2)}, 0, 0\right)^T$, där $\beta^{(1)} \equiv \iota$ betecknar interceptet och **X** har sex förklarande variabler.

- Definiera $\left(\hat{\beta}_r + \hat{\beta}'\right)$ som parametervekorn $\hat{\beta}_r$ utvidgad med en univariat parameterskattning $\hat{\beta}'$. Exempel: om $\hat{\beta}_4 \equiv \left(\hat{\beta}^{(1)}, \hat{\beta}^{(4)}, 0, \hat{\beta}^{(2)}, 0, 0\right)^T$, och $\hat{\beta}'$ skattar variabel nummer 3 i **X**, så har vi att $\left(\hat{\beta}_4 + \hat{\beta}'\right) \equiv \left(\hat{\beta}^{(1)}, \hat{\beta}^{(4)}, \hat{\beta}', \hat{\beta}^{(2)}, 0, 0\right)^T$.
- Definiera $\left(\hat{\beta}_r - \hat{\beta}^{(j)}\right)$ som parametervektorn $\hat{\beta}_r$ med sin komponent $\hat{\beta}^{(j)}$, $1 \le j \le r$ satt till noll. Exempel: $\left(\hat{\beta}_4 - \hat{\beta}^{(2)}\right) \equiv \left(\hat{\beta}^{(1)}, \hat{\beta}^{(4)}, 0, 0, 0, 0\right)^T$, $\hat{\beta}_4$ enligt ovan.
- Låt B_{-r} beteckna mängden av alla variabelnamn i **X**, som inte skattas i $\hat{\beta}_r$. *Exempel:* om **X** innehåller observationer för variablerna intercept, inkomst, utbildning och anställningstid, och $\hat{\beta}_r$ innehåller skattningar för intercept och anställningstid är $B_{-r} \equiv \{\text{'inkomst', 'utbildning'}\}$,
- Låt $p_{b,a} \equiv \mathrm{P}\left(\chi^2\left(m_b - m_a\right) > -2\lambda_{a,b}\right)$, beteckna p-värdet för likelihood-ratiotestet i (20) mellan $\ln\left(L\left(\hat{\beta}_a\right)\right)$ och $\ln\left(L\left(\hat{\beta}_b\right)\right)$, där $\hat{\beta}_a \subset \hat{\beta}_b$ och m_a och m_b är antalet parametrar i $\hat{\beta}_a$ och $\hat{\beta}_b$.
- Inför två stoppkriteria: $\alpha \in (0,1)$ och $\delta \in [\alpha, 1)$

Intitialsteg: Låt *r*=1, $\hat{\beta}_1 \equiv (\iota, 0, ..., 0)^T$, där ι betecknar interceptet. Vi börjar således med att finna ML skattningen för interceptet.

Additionssteg - utvidga modellen med den mest signifikanta variabeln med p< α :

1. Beräkna $\hat{\beta}_{tmp} = \left(\hat{\beta}_r + \hat{\beta}'\right)$, och därefter $p_{tmp,r}$ för alla univariata skattningar $\hat{\beta}'$ till variablerna i B_{-r}.
2. Låt $\hat{\beta}_{add}$ vara den parameterskattning $\hat{\beta}_{tmp}$, som fick lägst $\hat{p}_{tmp,r}$. (21)
3. *Stoppkriterium:* Om

$$\hat{p}_{add,r} < \alpha, \quad (22)$$

sätt $\hat{\beta}_{r+1} \equiv \hat{\beta}_{add}$. Låt $\hat{\beta}_{r+1} \equiv \hat{\beta}_r$ annars. Låt $r \equiv r+1$.

Subtraktionssteg - reducera modellen med den minst signifikanta variabeln med p< δ :

1. Beräkna $\hat{\beta}_{tmp} = \left(\hat{\beta}_r - \hat{\beta}'\right)$, och därefter $p_{r,tmp}$ för alla univariata skattningar $\hat{\beta}'$ av komponenterna i $\hat{\beta}_r$.
2. Låt $\hat{\beta}_{del}$ vara den parameterskattning $\hat{\beta}_{tmp}$, som fick lägst $\hat{p}_{r,tmp}$. (23)
3. *Stoppkriterium:* Om

$$\hat{p}_{r,del} < \delta, \quad (24)$$

sätt $\hat{\beta}_{r+1} \equiv \hat{\beta}_{del}$. Låt $\hat{\beta}_{r+1} \equiv \hat{\beta}_r$ annars. Låt $r \equiv r+1$.

Algoritmen har konvergerat när inga fler variabler läggs till eller tas bort från modellen, dvs när: $\hat{\beta}_r = \hat{\beta}_{r-1} = \hat{\beta}_{r-2}$. (25)

***

Förenkling: Om alla parametrar i β har samma antal frihetsgrader, så gäller att den parameter som har högst $\hat{p}_{r+1,r}$, alltid har högst $\ln\left(L\left(\hat{\beta}_{r+1}\right)\right)$ (se (20)). I detta fall kan således (21) ersättas med:

Låt $\hat{\beta}_{add}$ vara den parameterskattning $\hat{\beta}_{tmp}$ som fick högst $\ln(L)$. (26)

(26) gäller analogt för $\hat{\beta}_{del}$ i (23).

SAS använder sig av ett annat test än likelihood-ratiotestet (20) för beräkning av $\hat{p}_{tmp,r}$ i additionssteget (21). Detta test kallas för scoretest och är baserat på derivator av log-likelihooden[7]. Testet i subtraktionssteget (23) är baserat på Wald-testet (19).

Hosmer och Lemeshow (2000) nämner att likelihood-ratiotestet för (22) och (24) har bättre statistiska egenskaper och är att föredra i stepwisealgoritmen framför testen i SAS. Vi kommer att jämföra stepwisealgoritmen i SAS mot den likelihood-ratio baserade stepwisealgoritmen.

Hosmer och Lemeshow (2000) rekommenderar vidare 0.15-0.20 som lämpligt värde på α och δ. Vi kommer att diskutera denna rekommendation i avsnittet Informationskriteria och stepwisealgoritmen. I kapitlet Korsvalidering presenterar vi ett alternativt stoppkriterium.

Låt oss fortsätta med att presentera ett alternativ till likelihood-ratiotestet (20), som belyser svårigheten i att välja lämpliga värden på α och δ i stepwisealgoritmen.

Informationskriteria

När vi söker finna den bästa modellen, kan det vara användbart att ha ett mått på hur bra ett antal modeller är, utan att kräva att de är hierarkiskt ordade, begränsningar som det likelihood-ratio baserade testet (20) medför. Vi kallar sådana mått för informationskriteria.

Vi kommer i nästa avsnitt visa hur informationskriteria är relevanta för likelihood-ratiotestet.

Akaikes informationskriterium (AIC) är det mest kända informationskriteriet. AIC söker skatta $\ln\left(L\left(\beta|\boldsymbol{\mu}\right)\right)$, $\boldsymbol{\mu} \equiv \mathrm{E}[\mathbf{Y}]$, istället för $\ln\left(L\left(\beta|\boldsymbol{y}\right)\right)$, i (10).

[7] Se SAS Institute Inc (1999), kapitel 39.

För den logistiska regressionsmodellen gäller således att AIC försöker skatta $\ln\left(L\left(\beta|\mathbf{P}\right)\right)$ i (16), $\mathbf{P} \equiv \left(p_1,...,p_n\right)^T$.

Man kan visa att $\ln\left(L\left(\beta|\boldsymbol{y}\right)\right)$ växer med c:a 0.5 för varje variabel som läggs till i modellen. Om vi skulle använda oss av $\ln\left(L\left(\beta|\boldsymbol{y}\right)\right)$ som modellvalskriterium, skulle vi ta med alla variabler i modellen.

Man kan visa att för stora *n* gäller att

$$\widehat{\ln\left(L\left(\beta|\boldsymbol{\mu}\right)\right)} \approx \ln\left(L\left(\hat{\beta}|\boldsymbol{y}\right)\right) - k + \eta, \quad (27)$$

där *k* är antalet parametrar i modellen och η är en konstant oberoende av parametrarna i den generaliserade linjära modellen (6) för ett givet **X**, och sålunda redundant vid modelljämförelse. AIC definieras som (27) utan termen η multiplicerad med faktorn –2 (jfr. (20)):

$$AIC \equiv -2\ln\left(L\left(\hat{\beta}|\boldsymbol{y}\right)\right) + 2k \quad (28)$$

Modeller med lägst AIC bör således föredras vid modellurval då de maximerar $\ln\left(L\left(\hat{\beta},\boldsymbol{\mu}\right)\right)$. Termen 2*k* i (28) innebär att en modell *N* med *l* fler parametrar än en modell *M* kan anses vara bättre, endast om dess log-likelihood är *l* enheter större än log-likelihooden för M.

AIC har dock en tendens att överskatta antalet parametrar i modellen, eftersom den sanna modellen vanligtvis inte finns bland kandidatmodellerna, och det därför krävs fler variabler än nödvändigt för att skatta $\boldsymbol{\mu}$. Denna överskattning leder till att brus modelleras.

Andra kriterier som inte överskattar antalet parametrar i modellen är Bayesian Information Criterion (BIC) och Hannan-Quinn Information Criterion (HIC):

$$BIC \equiv -2\ln\left(L\left(\hat{\beta}\right)\right) + k\ln\left(n\right),\ HIC \equiv -2\ln\left(L\left(\hat{\beta}\right)\right) + 2k\ln\left(\ln\left(n\right)\right), \quad (29)$$

där *n* är antalet observationer.

BIC tenderar dock att välja för få parametrar i modellen. HIC är ett mellanting mellan AIC och BIC, som dock är något godtyckligt vald. Alla tre informationskriteria ovan kan skrivas på formen:

$$IC \equiv -2\ln\left(L\left(\hat{\beta}\right)\right) + k\varphi\left(n\right) \quad (30)$$

Informationskriteria i stepwisealgoritmen

I detta avsnitt visar vi att stepwisealgoritmen baserad på informationskriteria (30) är ekvivalent med stepwisealgoritmen baserad på likelihood-ratiotestet, om alla variabler i **X** har samma antal frihetsgrader.

Vi motiverar att om $\mathrm{P}\left[\chi^2(1) > \varphi(n)\right] = \alpha$, så gäller att $IC_N < IC_M \Leftrightarrow p_{N,M} < \alpha$, för två modeller *M* och *N*, med *k* respektive *k*+1 frihetsgrader.

Antag att vi endast har parametrar med en frihetsgrad i vår modellmatris **X** i stepwisealgoritmen. Då kan vi ersätta likelihood-ratiotestet (21) och (23) med log-likelihooden (26). Vi ser att (26) inte bara väljer modellen med lägst p-värde, utan även väljer den parameter som har lägst informationskriterium, eftersom *k* och *n* i (30) är lika för alla kandidatmodeller i additions- eller subtraktionssteget.

För att visa att stepwisealgoritmen med informationskriteria är ekvivalent med stepwisealgoritmen med likelihood-ratiotestet behöver vi även definiera ett stoppkriterium för varje α och δ i stepwisealgoritmen. För stepwisealgoritmen med informationskriterium på formen (30), är stoppkriteria (22) och (24) utbytta mot

$$-2\ln\left(L\left(\hat{\beta}_{r+1}\right)\right) + (k+1)\varphi(n) < -2\ln\left(L\left(\hat{\beta}_r\right)\right) + k\varphi(n)$$
$$\Leftrightarrow -2\ln\left(L\left(\hat{\beta}_{r+1}\right)\right) + \varphi(n) < -2\ln\left(L\left(\hat{\beta}_r\right)\right). \quad (31)$$

För kriterierna (22) och (24) gäller att:

$$\hat{p}_{r+1,r} < \alpha \Leftrightarrow \mathrm{P}\left[\chi^2(1) > -2\lambda_{r,r+1}\right] < \alpha \Leftrightarrow \mathrm{P}\left[\chi^2(1) > 2\left[\ln\left(L\left(\hat{\beta}_{r+1}\right)\right) - \ln\left(L\left(\hat{\beta}_r\right)\right)\right]\right] < \alpha \Leftrightarrow$$
$$2\left[\ln\left(L\left(\hat{\beta}_{r+1}\right)\right) - \ln\left(L\left(\hat{\beta}_r\right)\right)\right] > \mathrm{P}^{-1}_{\chi^2(1)}(\alpha) \Leftrightarrow -2\ln\left(L\left(\hat{\beta}_{r+1}\right)\right) + \mathrm{P}^{-1}_{\chi^2(1)}(\alpha) < -2\ln\left(L\left(\hat{\beta}_r\right)\right), \quad (32),$$

där $\mathrm{P}^{-1}_{\chi^2(1)}(\alpha) = x$ är inversen till fördelningsfunktionen $\mathrm{P}\left[\chi^2(1) > x\right] = \alpha$.

Jämför sista ekvationen i (31) och (32). Vi ser att kriterierna (31) och (32) är ekvivalenta när $\varphi(n) = \mathrm{P}^{-1}_{\chi^2(1)}(\alpha) \Leftrightarrow \mathrm{P}\left[\chi^2(1) > \varphi(n)\right] = \alpha$.

Vi kan exempelvis beräkna α för AIC, där $\varphi(n)$=2. Vi får att α_{AIC}=0.1573. (33)

Resultatet, att α_{AIC}=0.1573 tyder på att Hosmers och Lemeshows rekommendation med 0.15-0.20 som lämpligt värde på α (Hosmer och Lemeshow, 2000), kan ge upphov till modeller med för många parametrar i likhet med AIC-baserat modellval.

Diskussionen ovan kan samanfattas i följande resultat:

$$\text{Om } \mathrm{P}\left[\chi^2(1) > \varphi(n)\right] = \alpha, \text{ så gäller att } IC_N < IC_M \Leftrightarrow p_{N,M} < \alpha,$$

där

1. M och N, $M \subset N$, är två generaliserade linjära modeller med k respektive k+1 frihetsgrader, där informationskriteria $IC_M \equiv -2\ln\left(L\left(\hat{\beta}\right)\right) + k\varphi(n)$ och $IC_N \equiv -2\ln\left(L\left(\hat{\beta}\right)\right) + (k+1)\varphi(n)$, och
2. $p_{N,M} \equiv \mathrm{P}\left(\chi^2(1) > -2\left[\ln\left(L\left(\hat{\beta}_0\right)\right) - \ln\left(L\left(\hat{\beta}_1\right)\right)\right]\right)$, betecknar p-värdet för likelihood-ratiotestet i (20).[8]

Vi har motiverat att modellurval baserat på informationskriteria och likelihood-ratiotestet är ekvivalenta för parametrar med en frihetsgrad.

Vi nämnde i föregående avsnitt att informationskriteria riskerar att välja antingen för stora eller för små modeller. I nästa kapitel diskuterar vi modellvalidering och beskriver hur korsvalidering kan användas som stoppkriterium i stepwisealgoritmen.

Modellvalidering

Validering utför vi när vi utifrån en utvecklad modell med $\hat{\beta}$, skattar p_i på ett dataset, som inte använts för att utveckla modellen, och därefter beräknar ett mått en modells prediktiva styrka från $\hat{p}_i$. Modellvalidering utgör en kontroll på att vi inte överanpassat modellen på träningsdata.

Vid modellvalideringen kommer vi kommer att använda oss ett mått kallat för Accuracy Ratio (AR). AR verkar är en linjärkombination av det mer kända måttet ”Arean under ROC-kurvan”[9]. Ett linjärt samband gäller även mellan AR och Gini-indexet. AR och används i företaget för vilket scorekortet i Del III utvecklades.

AR får vi från ”liftkurvan”, som är en plot av

- % alla observationer ordnade efter fallande $\hat{p}_i$. Exempelvis motsvarar ”15%” på denna skala de 5% av alla observationer som har högst $\hat{p}_i$

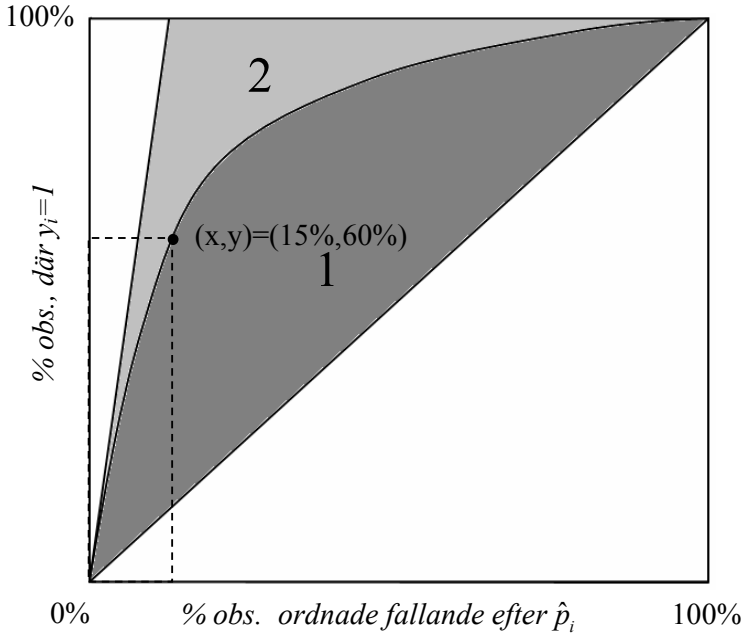
mot

- % av observationerna med $y_i = 1$. Exempelvis motsvarar 60% på denna skala 60% av alla observationer med $y_i = 1$.

[8] Resultatet ovan kan generaliseras till att gälla för fallet då modellen N har $k+l$ frihetsgrader. I det fallet byter vi ut $\chi^2(1)$ mot $\chi^2(l)$ och $\varphi(n)$ mot $l\varphi(n)$.

[9] Se Hosmer och Lemeshow (2000).

Figur 2. Areorna över och under liftkurvan.



Liftkurvan är kurvan mellan områdena 1 och 2 i Figur 2. Denna kurva mäter hur väl modellen separerar populationen där $y_i = 1$ från populationen där $y_i = 0$. Ju mer ”uppblåst” liftkurvan är över diagonalen i Figur 2, desto bättre är modellen på att separera $y_i = 1$ från $y_i = 0$. Arean under området 1 utgör arean under liftkurvan för den modell vi vill validera.

Diagonalen representerar den slumpmässiga modellen, där $\hat{p}_i$ är oberoende av y_i.

För den perfekta modellen gäller att $\hat{p}_i$ för alla observationer med $y_i = 0$ är lägre än $\hat{p}_i$ för alla observationer med $y_i = 1$. M.a.o. separerar den perfekta modellen populationerna där $y_i = 0$ och $y_i = 1$ fullständigt. I Figur 2 ger den mörkgrå arean 1 och den ljusgrå arean 2 tillsammans arean för den perfekta modellen.

AR jämför arean mellan liftkurvan och diagonalen med arean mellan liftkurvan för den perfekta modellen och diagonalen:

$$\text{AR} = \frac{1}{1+2}. \quad (34)$$

De streckade linjerna i Figur 2 visar att om en beslutsregel införs, som klassificerar 15% av populationen med högst $\hat{p}_i$ som $\hat{y}_i = 1$, så klassificerar vi korrekt ca 60% av populationen med $y_i = 1$. Sensitiviteten givet vår modell och beslutsregel är 60%.

Korsvalidering

En tredje metod för modellval istället för likelihood-ratiotestet eller informationskriteria kallas för korsvalidering.

Korsvalidering utgår från modellvalidering. En modell utvecklas på ett dataset och valideras den genom att beräkna ett valideringsmått, som AR (34) för ett annat dataset. Vid korsvalidering sker validering upprepade gånger för samma modell.

Metoden är den följande: Vi delar först upp träningsdata **X** i *k* delar. *k*-1 delar används som träningsdata för modellutveckling, och den återstående delen används som valideringsdataset, för vilket vi beräknar AR.

Denna process upprepas *k* gånger, så att alla *k* delar har använts som valideringdata. Medelvärdet för AR utgör sedan ett mått på modellens styrka.

Kohavi (1995) rekommenderar att välja *k* mellan 10 och 20 och att kontrollera urvalet av valideringsdata så att responsvariabeln innehåller samma proportion respons och icke-respons som träningsdata. Vi kommer i Del III att följa denna rekommendation och använda oss av *k*=10 och kontrollera urvalet m.a.p. responsvariabeln.

Korsvalidering och stepwisealgoritmen

Korsvalidering kan användas som stoppkriterium i stepwisealgoritmen. I stället för att bestämma α och δ eller något informationskriterium i (22) och (24), bestäms medel-AR i varje steg genom tiofaldig korsvalidering. Vi väljer sedan det första eller andra lokala maximumet.

Med korsvalidering i stepwisealgoritmen behöver vi således inte godtyckligt bestämma ett värde på α och δ, (22) och (24) i stepwisealgoritmen ersätts med $\mathrm{AR}_{add} > \mathrm{AR}_r$ och $\mathrm{AR}_{del} > \mathrm{AR}_r$.

Residualer [10]

Det är viktigt att granska residualerna i en modell för att kunna urskilja eventuell information som modellen inte lyckas förklara. Från (9) ovan har vi att $\mathrm{Var}(Y_i) = p_i(1-p_i)$, vi får således att residualen $\varepsilon \equiv Y_i - p_i$ har variansen $p_i(1-p_i)$. Vi ser att variansen varierar med p_i, med maximum i p_i=0.5. Vi väljer att standardisera variansen hos residualerna för att lättare kunna studera deras beteende.

Pearsonresidualen definieras som:

$$\hat{\varepsilon}_i^P \equiv \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}},$$

så att $\hat{\varepsilon}_i^P$ approximativt har väntevärde 0 och varians 1.

[10] Detta kapitel följer SAS Institute Inc. (1999), kapitel 39, sid.1963-1966, som följer, Pregibon, D. (1981): Logistic regression diagnostics, Annals of Statistics, Vol. 9.

Vi definierar vidare hattmatrisen **H** och $\hat{\mathbf{P}} \equiv \left(\hat{p}_1, ..., \hat{p}_n\right)^T$, där $\hat{\mathbf{P}} = \mathbf{HY}$. Diagonalelementet av **H** kallas för leverage och definieras som

$$\hat{h}_{ii} \equiv \widehat{\mathrm{Var}}\left(Y_i\right)\mathbf{X}_i\,\widehat{\mathrm{Var}}\left(\hat{\beta}\right)\mathbf{X}_i^T = \hat{p}_i\left(1-\hat{p}_i\right)\mathbf{X}_i\left(\mathbf{X}^T\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}_i^T,$$

där $\mathbf{X}_i$ är rad *i* i modellmatrisen **X**. Likheten ovan följer från (9) och (19). Leverage kan sägas uttrycka det potentiella inflytande på $\hat{p}_i$, som en observation kan ha.

Vi definierar Cook´s avstånd som

$$\Delta\hat{\beta}_i \equiv \left(\hat{\varepsilon}_i^P\right)^2\left[\frac{\hat{h}_{ii}}{\left(1-\hat{h}_{ii}\right)^2}\right].$$

Det kan visas att $\Delta\hat{\beta}_i$ mäter hur mycket parametervektorn β, standardiserad med kovariansmatrisen från (19), förändras om observation *i* tas bort. En annan tolkning av $\Delta\hat{\beta}_i$ är att $\Delta\hat{\beta}_i$ uttrycker den totala förändringen av $\mathrm{logit}\left(\hat{p}\right)$, nät observation *i* utesluts från beräkningen av $\hat{\beta}$.

Vidare har vi följande hjälpstatistika:

$$\Delta X_i^2 \equiv \frac{\left(\hat{\varepsilon}_i^P\right)^2}{1-\hat{h}_{ii}},$$

där ΔX_i^2 mäter bidraget av observation *i* till Pearson´s X^2 statistika, $X^2 \equiv \sum_{i=1}^{n}\left(\varepsilon_i^P\right)^2$.

Om Cook´s avstånd eller ΔX_i^2 är högt för någon observation *i*, är det värt att studera vilken av parametrarna i $\hat{\beta}$ som orsakar det höga värdet. Påverkan på parameter *l* i $\hat{\beta} \equiv \left(\hat{\beta}_1, ..., \hat{\beta}_k\right)$, $1 \le l \le k$ definieras som:

$$\Delta_l\hat{\beta}_i \equiv \frac{y_i - \hat{p}_i}{1-h_{ii}}\widehat{\mathrm{Var}_l}\left(\hat{\beta}\right)\mathbf{X}_i^T,$$

där $\widehat{\mathrm{Var}_l}\left(\hat{\beta}\right)$ är variansskattningen för parameter *l* i $\hat{\beta}$. $\Delta_l\hat{\beta}_i$ mäter hur mycket parameter *l* i $\hat{\beta}$ förändras om observation *i* tas bort.

Förbehandling av kategorialvariabler

Två begrepp från informationsteorin; ”Weight of Evidence” (WoE) och Informationsvärde (IV), används för att reducera antalet kategorier hos kategorialvariabler.[11] Vi redogör i korthet för dessa begrepp.

”Weight of Evidence” definieras som:

$$\mathrm{WoE}_{i,j}(x) \equiv \ln\left(\frac{P(x|y=i)}{P(x|y=j)}\right). \quad (35)$$

I WoE-funktionen figurerar den förklarande variabeln *x* som en stokastisk variabel, vilket gör WoE till en Bayesiansk funktion. Vi kan visa att denna funktion har sin motsvarighet i den logistiska regressionsmodellen:

$$\mathrm{WoE}_{1,0}(x) \equiv \ln\left(\frac{P(x|y=1)}{P(x|y=0)}\right) = \ln\left(\frac{P(y=1|x)P(x)P(y=0)}{P(y=0|x)P(x)P(y=1)}\right) =$$
$$= \ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) - \ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta x - \iota,$$

där βx är en logistisk regressionsmodell (4) för en stokastisk variabel *x,* och ι är en logistisk regressionsmodell med enbart intercept. WoE mäter således hur mycket βx skiljer sig från modellen med endast intercept.

Vi definierar Kullback-Leibleravståndet som:

$$\mathrm{D}_{(x|y=i),j} \equiv \mathrm{E}_{x|y=i}\left[\mathrm{WoE}_{i,j}(x)\right].$$

$\mathrm{D}_{(x|y=1),0}$ mäter hur mycket väntevärdet av den logistiska regressionsmodellen för *x,* givet *y*=1, avviker från modellen med endast intercept. $\mathrm{D}_{(x|y=1),0}$, kan ses som ett mått på den extra information som variabeln *x* tillför modellen när y=1. Om vi adderar $\mathrm{D}_{(x|y=1),0}$ och $\mathrm{D}_{(x|y=0),1}$ får vi ett mått på den extra information som *x* tillför modellen totalt. Detta mått kallas för informationsvärdet (IV). Märk att IV alltid är positivt och ökar ju större skillnaderna är mellan fördelningarna $P(x|y=1)$ och $P(x|y=0)$:

$$\mathrm{IV} \equiv \mathrm{D}_{(x|y=1),0} + \mathrm{D}_{(x|y=0),1} = \sum_x P(x|y=1)\,\mathrm{WoE}_{1,0}(x) - \sum_x P(x|y=0)\,\mathrm{WoE}_{1,0}(x) =$$
$$= \sum_x \left(P(x|y=1) - P(x|y=0)\right)(\beta x - \iota), \quad (36)$$

[11] Förfarandet rekommenderas i Siddiqi (2006) och i företagets interna scorekortsprocesser.

där vi i första likheten betingat väntevärdet på y samt utnyttjat att $\mathrm{WoE}_{0,1}(x) = -\mathrm{WoE}_{1,0}(x)$. IV kan skattas från (36) ovan, genom $\hat{p}_j(i) \equiv \hat{P}(x = i | y = j) = n_{x=i,y=j} / n_{y=j}$, där $\hat{p}_j(i)$ är en ML-skattning när vi har $n_{x=i,y=j}$ lyckade av totalt $n_{y=j}$ Bernoulliförsök.

WoE-kodning

Ibland kan det vara önskvärt att transformera kategorialvariabler till intervallvariabler, genom en process kallad WoE-kodning. WoE-kodning innebär att varje variabelvärde ersätts med WoE-värdet för den kategori, som värdet tillhör.

Fördelen med WoE-kodningen, är att den reducerar antalet frihetsgrader för variabeln till 1, precis som för intervallvariabler. Å andra sidan försvinner möjligheten att separat anpassa koefficienterna för variabelns kategorier i den logistiska regressionsmodellen.

Del III. Modelleringsexempel

I Del III, redovisar vi hur begreppen från Del II kan användas vid kreditvärdering. Vi börjar med att i kapitlen Databeredning och Förbehandling av variabler diskutera redovisa hur vi konstruerat vår datamatris med förklarande variabler **X** och vår responsvariabel i **Y**. Olika formuleringar av stepwisealgoritmen jämförs.

Våra data innehåller information om ett antal personlån. De utvecklade logistiska regressionsmodellerna söker förutsäga sannolikheten för att ett nytt lån inte kommer betalats tillbaka vid ansökningstillfället.

Databeredning

Detta avsnitt beskriver hur data samlats in och förberetts för statistisk analys. Vidare definieras responsvariabeln för den logistiska regressionsmodellen samt kriteria för att utesluta data från modelleringsprocessen.

Datakarakteristik

Vår datamängd innehåller data för ett kontantlån från ett finansföretag insamlade från lanseringen av lånet den 26.4.2005 till den 20.5.2006. Under denna tid registrerades 3914 lån. Dessa lån utgör vår lånedatabas.

Observationsdatum för att ”läsa av” betalningshistoriken för lånen är 20.5.2006. Lånen har i medeltal 7.8 föreskrivna amorteringar vid observationsdatumet. Maximala och minimala antalet föreskrivna amorteringar är 1 respektive 12. Lånetiden varierar mellan 12 och 36 månader.

Variablerna i datasetet kan delas in i två typer: ansökningsvariabler och beteendevariabler.

Ansökningsvariabler är parametrar som samlats in när klienten ansökte om sitt lån. Exempel på ansökningsdata är:

- Yrke
- Boende
- Inkomst
- Tidpunkt för ansökan
- Föremål som finansieras

Beteendevariabler innehåller information som beskriver klientens beteende vid tidigare lån. Beteendedata har ofta större förmåga att förutsäga betalningsförmåga än ansökningsdata.

Exempel på beteendedata är:

1. Antal lån som avslutats p.g.a. betalningsförsummelse

2. Antal amorteringar som klienten ålagts att betala, men som vid observationstillfället är obetalda.

Datasetet innehåller

- 102 ansökningsvariabler
- 9 beteendevariabler för tidigare lån
- 4 beteendevariabler för det aktuella lånet:
    1. Lånefas. Anger om lånet är aktivt eller avslutat samt orsaken till att lånet är avslutat, t. ex. avlutat p.g.a. betalningsförsummelse.
    2. Antal betalda och obetalda föreskrivna amorteringar med föreskrivet betalningsdatum innan observationsdatum.
    3. Datum för första föreskrivna amorteringen.

Utöver lånedatabasen har vi en andra databas, som omfattar alla amorteringar för lånedatabasen med förfallodag fram till observationsdatumet 20.5.2006. Amorteringsdatabasen innehåller följande variabler:

1. Kontraktnummer
2. Ordningstal för amorteringen
3. Förfallodag
4. Betalningsdag
5. Föreskrivet amorteringsbelopp
6. Belopp betalt av klienten

Beviljningsprocess

Beviljningsprocessen för lånen, fungerar enligt följande:

1. Kunden ringer in till företaget efter att ha sett en annons i tidningen eller på TV, eller hört reklamen på radio. Kunden får svara på ett antal frågor, som namn, adress, personnummer osv.
2. Kunden blir ”scorad” med hjälp ett generiskt scorekort. Scorekortet är ej utvecklat för ett kontantlån, utan för en annan produkt.
3. Utöver scorekortet, finns också sk. knock-out regler, som direkt avvisar klienten. Exempel på knock-out regler är:
   Avvisa låneansökan om
    a. klienten har ett försummat lån i ett externt kreditregister, där skulden ej är återbetald. Ett lån betecknas som försummat om klienten försummat att betala tre amorteringar i följd.
    b. klienten är studerande, värnpliktig eller arbetslös.
4. Om kunden får sitt lån beviljat skickar företaget ett brev med lånedokumentation, samt en lista med andra dokument som klienten måste skicka in. Exempel på sådana dokument är inkomstintyg från arbetsgivaren och de två sista kontoutdragen.
5. Om klienten skickar tillbaka de begärda dokumenten kontrolleras hans uppgifter genom att ringa upp klienten och hans arbetsgivare samt genom att jämföra de bifogade dokumenten med de uppgifter klienten delgav per telefon.
6. Om kontrollen av klienten inte ger skäl till att avvisa låneansökan, betalas pengarna ut på klientens bankkonto.

Endast de låneansökningar som blir beviljade efter steg sex ovan ingår i vår datamängd. En enskild kund kan få maximalt ett lån beviljat.[12]

Tillvägagångssätt vid betalningsförsummelse

I finansföretaget tillåts klienten att ha maximalt tre obetalda amorteringar vid ett enskilt tillfälle. En obetald amortering definierar vi en amortering som är obetald vid slutet av månaden för sin förfallodag, som infaller den 15. i månaden.

Om fyra amorteringar samtidigt är obetalda, avslutas kontraktet och klienten blir skyldig att betala tillbaka de obetalda amorteringarna plus det resterande framtida lånebeloppet (exklusive ränta och de obetalda amorteringarna). Ärendet går till inkassofirmor, eller polisen om bedrägeri misstänks föreligga. Om klienten efter en längre tid inte betalat tillbaka sin skuld går ärendet till domstol.

Definition av ett bra, dåligt samt obestämt kontrakt

Vi vill förutsäga lån där företaget går med förlust, vanligtvis lån som inte betalas tillbaka. Låt oss kalla ett lån inte kommer betalas tillbaka för ett dåligt lån.

Från lån som avlutats p.g.a. betalningsförsummelse inkasseras endast ca 20% av det inkrävda beloppet. Detta från erfarenhet som låneföretaget har från andra produkter. Det verkar vara en vettig definition att definiera ett dåligt kontrakt som ett kontrakt som har avslutats p.g.a. betalningsförsummelse.

Följdfrågan som man kan ställa sig är om inte kontrakt som har två eller tre försummade amorteringar i framtiden löper lika stor risk att förfalla.

Vi undersökte hur ofta ett lån är avslutat åtta månader efter förfallodagen om det har x obetalda amorteringar y månader efter förfallodagen för den första avbetalningen, $y<8$.

Vi valde inte någon längre observationsperiod än åtta månader eftersom de äldsta kontrakten i vår låneportfölj inte än hade sin tolfte avbetalning föreskriven innan observationsdatumet 20.6.2006. En längre observationsperiod skulle resulterat i få antal kontrakt att analysera.

Vi valde omvänt inte en kortare observationsperiod än åtta månader för att ge ett lån möjlighet att avslutas från att ha haft x obetalda amorteringar y månader efter förfallodagen för den första avbetalningen.

[12] En analys kallad för reject inference utförs ibland, vilken innebär att man inkluderar de ansökningar som avvisats av det tidigare scorekortet i modelleringsprocessen, genom att anta något värde på responsvariabeln. Se Siddiqi 2006 för introduktion.

Tabell 2. Andel lån avslutade för betalningsförsummelse åtta månader efter förfallodagen för första amorteringen, uppdelade efter antalet obetalda amorteringar (*x-axel*) 3-6 månader efter förfallodagen för första amorteringen (*y-axel*). 2111 kontrakt, som hade betalningsdagen för sin första amortering åtta månader bakåt i tiden eller mer ingår.

	Antal obetalda amorteringar			
Månader efter förfallodagen för första amorteringen	**0**	**1**	**2**	**>=3**
3	3.3%	24.5%	66.7%	73.5%
4	1.4%	22.9%	56.3%	86.3%
5	0.0%	15.0%	50.0%	82.6%
6	0.0%	0.0%	44.0%	76.3%

Ett och samma kontrakt ingår i Tabell 2 för alla rader i tabellen, t.ex. kan ett kontrakt ha 1 obetald amortering 3 månader efter förfallodagen, och sedan 0 obetalda 4 månader efter förfallodagen osv.

Vi ser från Tabell 2 ovan att c:a 80% av lånen med tre obetalda amorteringar är avskrivna åtta månader efter dagen för första amorteringen. C:a 50% av lånen med två obetalda amorteringar är avskrivna vid åttamånadersgränsen. Av lånen med en obetald amortering är c:a 20% avskrivna. Motsvarande siffra för lån utan obetald amortering är ca 3%.

Märk att ett lån med 0 obetalda amorteringar 5 månader efter första avbetalningsdatumet aldrig kan vara avskrivet åtta månader efter första avbetalningsdatumet. Maximalt kan lånet ha tre obetalda amorteringar. Som vi nämnde ovan krävs fyra obetalda amorteringar för att lånet ska avslutas p.g.a. betalningsförsummelse.

Efter expertbedömning av informationen ovan och med erfarenhet från tidigare scorekort, kom vi fram till följande definition av vår observerade responsvariabel för kontrakt nummer i (jfr. (4)):

$$y := \begin{cases} 0 \text{ , om 0 obetalda amorteringar och } \geq \text{tre amorteringar betalda} \\ 1 \text{ , om } \geq 3 \text{ obetalda amorteringar eller avslutat lån p.g.a. betalningsförsummelse} \end{cases} \tag{37}$$

Det bedömdes som osäkert om kontrakt med 1 eller 2 obetalda kommer ge upphov till en vinst eller förlust. Dessa kontrakt uteslöts ur datamängden för att undvika ökad osäkerhet i parameterskattningarna. Ovanstående praxis att utesluta osäkra data beskrivs närmre i Siddiqi (2006) och Thomas (2000).

Vidare bestämdes att kontrakt som har 0 obetalda innan tre avbetalningar är betalda ej med säkerhet kan gå med vinst. Normalt utesluts kontrakt som ej har tolv föreskrivna avbetalningar. Detta för att man skall vara säker på att ca 80-90% av lånen som kommer avslutats p.g.a. betalningsförsummelse under hela lånetiden avslutats vid denna tidpunkt.

Då alla lån i vår datamängd var ett år eller yngre, fick denna korta observationsperiod på tre månader eller mindre användas istället.

Mer avancerade definitioner på bra, dåliga och obestämda kontrakt övervägdes, innehållande mer information om fördröjningen med vilken amorteringarna betalades, t. ex.

maximalt obetalda sista sex månaderna. Då vi hade få data att jämföra med, och en ung låneportfölj, bedömdes tillförlitligheten av en mer finkornig definition som osäker.

I vår modelleringsprocess kommer vi således att söka skattningar på β, som förutsäger y i (37) ovan så bra som möjligt.

Tidiga observationer

De två första veckorna efter produktlanseringen justerades granskningsförfarandet i punkt 6. i beviljningsprocessen. Vidare antas att fler bedragare och personer i stora finansiella svårigheter än normalt, sökte och beviljades lån. Detta av orsaken att en ny aktör dök upp på marknaden för snabba telefonlån i och med produktlanseringen.

Av dessa orsaker betraktades urvalet de två första veckorna inte som representativt och dessa data uteslöts från urvalet.

Lån avslutade av andra anledningar än försummelse

Ett fåtal lån avslutades p.g.a. att klienten betalat tillbaka sitt lån i förtid. Vid tidig återbetalning försvinner den mesta räntevinsten. Dessa lån kunde varken sägas ge upphov till en vinst eller en förlust och uteslöts således.

Sammanfattning, uteslutna lån

Tabell 3. Orsaker till uteslutning från lånedatabasen med 3914 lån

Orsak till uteslutning	**% uteslutna**	**Antal**
Ej tre betalda avbetalningar och 0 obetalda amorteringar	11%	489
En eller två obetalda amorteringar	10%	376
Ansökt innan 15. maj 2005	16%	609
Avslutad av annan anledning än betalningsförsummelse	0.3%	13
Totalt	35%	1435

N.B. vissa lån kan vara uteslutna av mer än en orsak.

Antalet kvarvarande kontrakt delades upp i ett valideringsdataset och ett träningsdataset genom slumpmässigt urval, kontrollerat m.a.p. responsvariabeln och antalet återbetalda inbetalningar. 25% av observationerna användes som valideringsdata och 75% som träningsdata. Tabell 4 sammanfattar de tre databaserna:

Tabell 4. Databaser skapade från den initiala datamängden

Databas	Antal
Träningsdata	1859
Valideringsdata	620
Uteslutna	1435
Totalt	3914

Identifikation av felaktiga data

För alla parametrar granskades de högsta och lägsta värden för att hitta eventuella fel i datainskrivningen. Tre kontrakt identifierades som felaktigt inskrivna. Alla fel hänförde sig till variabeln ”anställningsår”. Anställningsår 1111 ersattes med 2006, 1194 med 1994 och 1199 med 1999.

Hantering av saknade data

Hos flera variabler saknades datavärden. Data kunde saknas av två principiella orsaker:

1. Klienten uppgav ingen information. T. ex. klienten mindes inte vilken månad han anställdes i sin nuvarande anställning.
2. Data var odefinierade. T. ex. om låntagaren saknade medlåntagare saknades värden på alla variabler gällande medlåntagaren.

Saknade data ersattes enligt följande:

- För kategorialvariabler skapades en ny kategori, kallad ”-1”.
- Det bedömdes att klienten inte alltid kunde antas minnas den exakta månaden för datum som låg mer än två år tillbaka i tiden. Dessa data registrerades ej.

  Saknade månadsdata fick värdet 7 (d.v.s. juli), om årtalet för samma datatyp ej saknades. Vi valde värdet 7, då medelvärdet av alla månader är 6,5. Detta värde antogs vara något så när väntevärdesriktigt, förutsatt att alla månader var lika representerade i datamängden.

  Variabler, där månad och år saknades, t.ex., anställningsmånad och år behandlas i kapitlet Förbehandling, avsnittet Intervallvariabler av variabler.
- Övriga ordinal- och intervallvariabler tilldelades värdet -1.
- En indikatorvariabel skapades för varje variabel som saknade värden. Indikatorvariabeln tilldelades värdet 1 när varabeln i fråga saknade värde och tilldelades värdet 0 i övriga fall. Förfarandet att skapa indikatorvariabler för saknade värden på en variabel rekommenderas av Hosmer och Lemeshow (2000) och har applicerats av Vogel, Gottschalk & Wang (2004).

Förbehandling av variabler

Vi behandlade kategorialvariabler och intervallvariabler separat. Alla variabler studerades med avseende på hur de påverkade risken för betalningsförsummelse.

Nya variabler som vi kunde anta ha inverkan på vår responsvariabel skapades. Vi skapade dels förklarande ”atomära” variabler, som anställningstid (ansökningsdatum minus anställningsdatum), dels samspelsvariabler för finansiella variabler, som disponibel inkomst efter lån (inkomster minus utgifter). Vi valde inte att pröva fler samspelsvariabler p.g.a. att vår datamängd var liten och den stora risken för att få med ett samspel som negativt påverkade Accuracy Ratio (34) vid validering.

Kategorialvariabler

I detta avsnitt beskriver vi hur vi använde oss av ”Weight of Evidence” (WoE) (35) och informationsvärde (IV) (36) för att minska vi antalet kategorier för varje kategorialvariabel[13].

Därefter går vi igenom hur vi valde att antingen WoE-koda eller dummy-koda kategorialvariabeln.

Sammanslagning av kategorier:

Variabler som omfattar kategorier, exempelvis civilstånd eller yrke är inte kvantifierbara. Kvantifierbarhet är nödvändig för att kunna beräkna $\mathbf{X}\beta$ i (4).

I Tabell 5 nedan visas ett exempel på hur informationsvärdet används för att studera kategorialvariabler. Vi söker att reducera antalet kategorier för att undvika att modellen överanpassas. Vi studerar framförallt de två kolumnerna längst till höger: ”andel av IV” och ”andel dåliga lån per kategori”. Det är viktigt att varje gruppering kan motiveras, så att inte artificiella kategorier skapas.

Tabell 5. Tabell för informationsvärde för variabeln ”Aktivitetstyp”.

Värde	*Antal*	*Antal goda lån*	*Antal dåliga lån*	*Andel av alla goda lån*	*Andel av alla dåliga lån*	*WoE*	*IV per kategori*	*Andel av IV*	*Andel dåliga lån per kategori*
	n	$n_{y=0}$	$n_{y=1}$	$\hat{p}_0(i)$	$\hat{p}_1(i)$				$n_{y=1}/n$
Pensionär	212	167	45	0.14	0.07	-0.62	0.0391	0.5432	0.21
Förtidspensionär	188	133	55	0.11	0.09	-0.20	0.0037	0.0519	0.29
Föräldraledig	33	21	12	0.02	0.02	0.13	0.0003	0.0041	0.36
Företagare	145	82	63	0.07	0.10	0.42	0.0148	0.2062	0.43
Statsanställd	287	202	85	0.16	0.14	-0.18	0.0047	0.0659	0.30
Privatanställd	994	632	362	0.51	0.58	0.13	0.0093	0.1287	0.36
TOTALT	1859	1237	622	1	1		0.0720	1.0000	

I exemplet ovan svarar kategorin Pensionär för mer än hälften av IV; andelen dåliga lån är överlägset lägst. Denna kategori låter vi stå för sig. Förtidspensionärer klumpas ihop med

[13] Detta förfarande beskrivs i Siddiqi (2006) och Hand (2005).

Statsanställda. Denna kategori kan kallas ”offentlig trygghet”. Märk att andelen dåliga lån för Förtidspensionärer och Statsanställda ligger nära varandra. Övriga kategorier slås ihop i en kategori, som vi väljer att kalla ”privat sektor”. Efter att de nya kategorierna skapats kontrollerar vi IV och ser att det sjunkit med 9% från 0.0720 till 0.0657. Ytterligare sammanslagning av kategorier minskar IV med mer än 30%.

En tumregel i Siddiqi (2006) klassificerar styrkan hos informationsvärdet enligt Tabell 6:

Tabell 6. Tumregel för IV-styrka.

IV för en variabel	Prediktiv styrka
<0.02	Ej prediktiv
0.02-0.1	Svag
0.1-0.3	Medelstark
>0.3	Stark

I vår datamängd hade den starkaste variabeln ett informationsvärde på 0.13. Bara två variabler hade IV> 0.1. Detta är en indikation på att vårt dataset inte hade stark prediktiv förmåga.

Ingen variabel i den finala modellen tilläts ha färre observationer per kategori än 8% av alla observationer, med undantag för tre kategorialvariabler. Två av dessa tre variabler hade en risknivå på mer än 50%, vilket motiverade att de togs med. Den sista variabeln togs med av affärsmässiga orsaker, då den beskrev klientens tidigare samröre med företaget.

Weight of Evidence-kodning:

För två modeller, vilka beskrivs i kapitlet Modellering nedan, valde vi att WoE-koda den nya kategorialvariabeln (se avsnittet WoE-kodning i kapitlet Förbehandling av kategorialvariabler i Del II). WoE-kodning innebär att variabeln transformeras från att vara en kategorialvariabel till en intervallvariabel, som antar samma värden som WoE för respektive kategori. Om vi skulle vilja WoE-koda vår nya aktivitetstyp med sammanslagna kategorier i exemplet ovan, skulle den nya variabeln värdet -0.62 för observationer med kategorin ”pensionär” (se kolumn WoE i Tabell 5).

Dummy--kodning:

För en modell i kapitlet Modellering valde inte att WoE-koda kategorialvariablerna, utan använde oss istället av dummykodning.

Dummykodning innebär att varje kategori utom en kodas som indikatorvariabel (eller dummyvariabel). Den kategori som utesluts från dummykodningen kallas för referensvariabel. Referensvariabeln kan konstrueras som en linjärkombination av de andra kategorierna. När variabelns p-värde utvärderas (jfr. (20)), får den samma antal frihetsgrader, som antalet indikatorvariabler, d.v.s. antalet kategorier minus ett. Varje indikatorvariabel tilldelas en motsvarande parameter i β.

Exempel på dummykodning: Antag att vi har en variabel kallad för ”högsta avslutad utbildning” med tre olika kategorier: ”grundskola”, ”gymnasium” och ”universitet”. Från denna variabel skapar vi två nya indikatorvariabler, v_1 och v_2, vilka fångar all information i variabeln ”avslutad utbildning”:

$$v_1 \equiv \begin{cases} 1, & \text{om avslutad utbildning="grundskola"} \\ 0, & \text{annars} \end{cases}$$

$$v_2 \equiv \begin{cases} 1, & \text{om avslutad utbildning="gymnasium"} \\ 0, & \text{annars} \end{cases}$$

Vi behöver inte separat konstruera en tredje variabel v_3 för värdet ”universitet”, då vi vet att om $v_1 = v_2 = 0$, så är ”högsta avslutad utbildning”=”universitet”. I den logistiska regressionsmodellen får variabeln ”högsta avslutad utbildning” två frihetsgrader. v_1 och v_2 ingår i **X**.

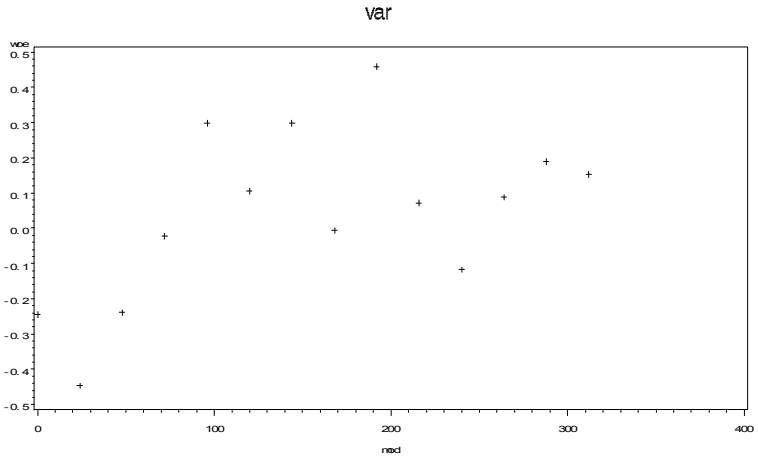
Intervallvariabler

Intervallvariabler analyserades med IV, genom att avrunda värdena, så att kategorier jämna intervall skapades (se Figur 3 nedan). Vi plottade sedan kategorierna av variabeln mot dess WoE och studerade om sambandet mellan de avrundade värdena och WoE var linjärt och om inte ett fåtal observationer riskerade att bli för inflytelserika.

Hos vissa variabler avrundades höga eller låga värden för att undvika inflytelserika outliers. I Figur 3 avrundades värden som var högre än 168 av till 168.

Indikatorvariabler skapades om det verkade som om andelen dåliga lån snabbt ändrades, eller om de flesta data saknades (t.ex. variabeln ålder för medlåntagaren). För variabeln i Figur 3 nedan skapades en indikatorvariabel som antog värdet 1 för värden < 66 och 0 annars.

Figur 3. Plot av en intervallvariabel.



Intervallvariabler med odefinierade värden (se kapitel Databeredning, avsnitt Hantering av saknade data), som saknades för mindre än 50% av data, ersattes det värde som motsvarade den risknivå som de saknade värdena hade.

Exempel: variabeln anställningstid saknades för klienter utan arbete (pensionärer, hemmafruar). Genom att avrunda anställningstiden kunde risknivån för varje kategori analyseras enligt ovan. De saknade värdena på anställningstiden tilldelades en anställningstid, som hade motsvarande risknivå, som kategorin med de saknade värdena.

Några intervallvariabler som inkomst transformerades med logaritmfunktionen. En variabel, som visade sig ha höga residualer (se avsnittet Studium av residualer nedan) delades upp i olika intervall och behandlades som kategorialvariabel.

Modellering

I vårt modellbygge prövades tre ansatser:

Ansats 1. Stepwisealgoritmen i SAS® med scoretestet i additionsteget och Wald-testet i subtraktionssteget (se kommentar under (26) ovan) och WoE-kodade kategorialvariabler. Förurval av variabler som kunde inkluderas i modellen utfördes baserat på p-värdet från enkel logistisk regression med WoE-kodade kategorialvariabler.

Ansats 2. Stepwisealgoritmen med likelihood-ratiotestet och dummykodade kategorialvariabler.

Ansats 3. Stepwisealgoritmen med likelihood-ratiotestet och WoE-kodning av kategorialvariabler.

För varje steg i stepwisealgoritmen beräknades medel-AR genom tiofaldig korsvalidering, där vi kontrollerade samplingen av valideringsdata med avseende på responsvariabeln. Den bästa modellen utgjorde det första lokala minimumet på medel-AR. I Ansats 1 och 3 WoE-kodades variablerna för varje korsvalidering på träningsdata utan tiondelen av träningsdata, som användes för validering. För alla ansatser satte vi $\alpha = \delta = 0.3$, för att vara säkra på att modellen med högst medel-AR fanns med i serien av utvecklade modeller (se avsnitt Korsvalidering och stepwisealgoritmen).

Korsvalidering och stepwisealgoritmen med likelihood-ratiotestet programmerades i SAS.

Nedan beskrivs de tre ansatserna och medelvärdet på AR (34) efter korsvalidering.

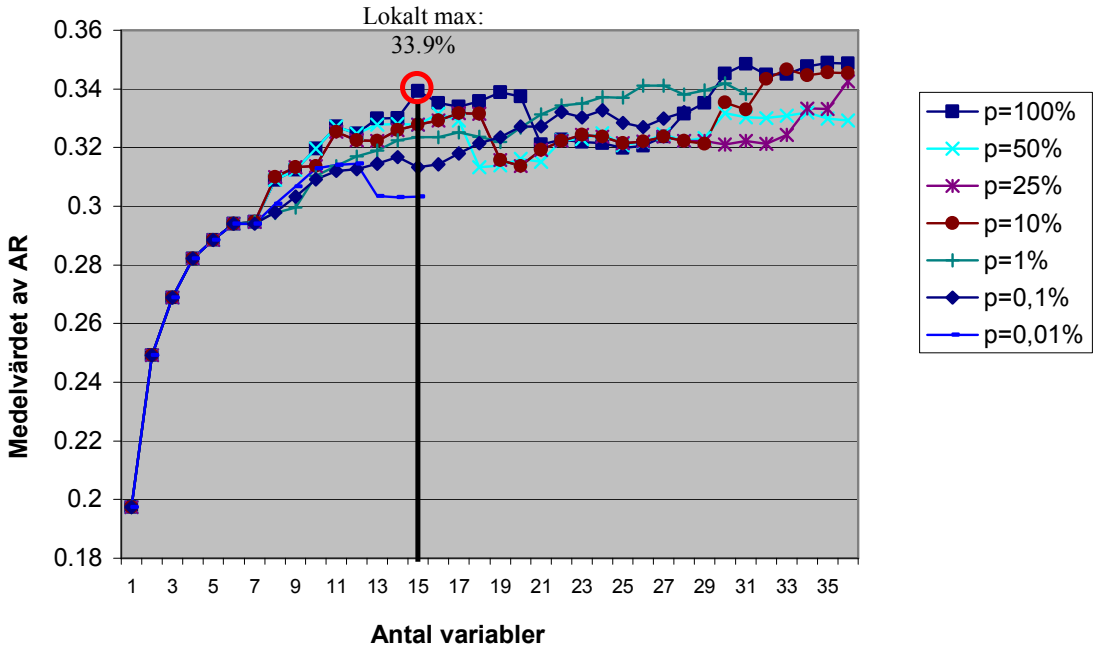
Modelleringsresultat:

Ansats 1:

Efter förbehandlingen av variablerna enligt föregående avsnitt hade vårt dataset mer än 100 variabler. Kategorialvariabler WoE-kodades, för att undvika att de diskriminerades för sina frihetsgrader. Enkel logistisk regression utfördes på varje variabel.

Från datasetet valde vi kandidatvariabler till stepwisealgoritmen, vilkas p-värde från den enkla logistiska regressionen var lägre än 0.01%, 0.1%, 1%, 10%, 25%, 50% och 100%. Ett p-värde innebar att alla variabler i datasetet togs med som kandidatvariabler till stepwisealgoritmen. Resultatet visas nedan.

Figur 4. *Ansats 1.* Medelvärdet för AR efter tiofaldig korsvalidering för varje steg i stepwisealgoritmen i SAS med scoretest i additionssteget och Wald-test i subtraktionssteget och WoE-kodade kategorialvariabler, för olika maximum på p-värden i enkel logistisk regression för kandidatvariabler.



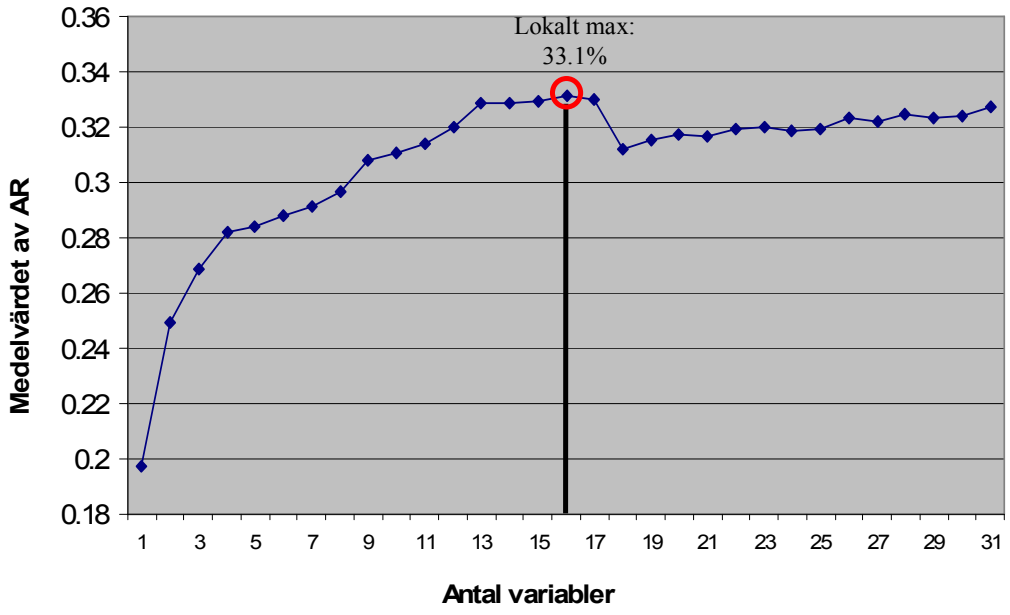
Vi ser från Figur 4 att förurval av kandidatvariabler till stepwisealgoritmen inte förbättrar medelvärdet på AR vid korsvalidering.

Även om vi uppnår högst värde på medel-AR för 31 variabler när p=100%, är det troligen en följd av överanpassning till träningsdata. Vi väljer således modellen med 15 variabler, som utvecklades utan att filtrera kandidatvariabler.

Ansats 2:

Stepwisealgoritmen med likelihood-ratiotestet. Vi dummy-kodade kategorialvariablerna. Medelvärdet på AR beräknades vid varje steg.

Figur 5. *Ansats 2.* Medelvärdet för AR vid tiofaldig korsvalidering för varje steg i stepwise algoritmen med likelihood-ratio test och med dummy-kodade variabler.



Från Figur 5 ser vi att den bästa modellen har 16 variabler.

Ansats 3:

Stepwisealgoritmen med likelihood-ratio testet. Vi WoE-kodade kategorialvariablerna så att de alla hade en frihetsgrad. Medelvärdet på AR beräknades vid varje steg.

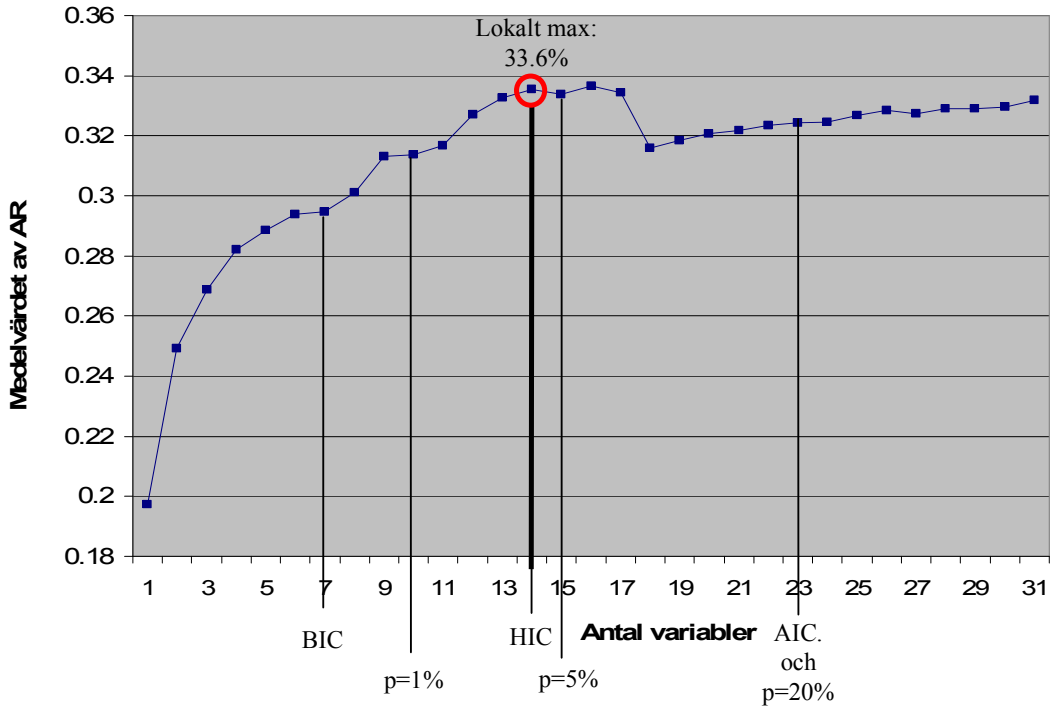
Vi beräknade även värdet på $\varphi(n)$ och $P^{-1}_{\chi^2(1)}(\alpha)$ enligt (31) och (32). Resultatet redovisas nedan i Tabell 7. Vi ser att BIC ligger nära α=0.01 och HIC nära α=0.05. Från (33) har vi att AIC motsvarar α=0.157, och således ligger nära α=0.2.

Tabell 7. Värdet på $\varphi(n)$ i (31), *n*=1859, för BIC, HIC och AIC samt värdet på $P^{-1}_{\chi^2(1)}(\alpha)$ i (32) för $\alpha=\delta$=0.01, 0.05 och 0.2.

Kriterium	Värde
$\varphi_{BIC}(1859)$	7.5
$\varphi_{HIC}(1859)$	4.0
$\varphi_{AIC}(1859)$	2
$P^{-1}_{\chi^2(1)}(0.01)$	6.6
$P^{-1}_{\chi^2(1)}(0.05)$	3.8
$P^{-1}_{\chi^2(1)}(0.2)$	1.6

Modellerna som valdes m.a.p ovanstående sex kriteria redovisas i Figur 6 nedan. Vår valda modell efter korsvalidering är indikerad med en tjock lodrät linje.

Figur 6. *Ansats 3.* Medelvärdet för AR vid korsvalidering för varje steg i stepwisealgoritmen med likelihood-ratiotest och WoE-kodade variabler; samt finala modeller efter stepwisealgoritmen för informationskriteria och p-värden som stoppkriteria.



Från Figur 6 ser vi att den bästa modellen har 14 variabler. Vi har ett andra maximum för en modell med 16 variabler, vilket dock är försumbart högre än det första lokala maximumet.

Vi ser att informationskriterierna BIC och AIC missar maximum. Det samma gäller för likelihood-ratio urvalet med $p=\alpha=\delta\in\{0.01,0.2\}$. HIC väljer rätt modell och $p=\alpha=\delta=0.05$ ligger nära.

Vi sammanfattar resultaten från Ansats 1, 2 och 3 i Tabell 8:

Tabell 8. Resultat från den bästa tiofaldigt korsvaliderade modellen från stepwisealgoritmen för Ansats 1, 2 och 3.

Ansats	Medel-AR på valideringsdata från korsvalidering för vald modell	Variabler för vald modell
1. SAS stepwise, WoE kodning	33.9%	15
2. Likelihood-ratio, dummy-kodning	33.1%	16
3. Likelihood-ratio, WoE kodning	33.6%	14

Vi kan se att alla modeller är ungefär lika starka. Stepwisealgoritmen i SAS verkar fungera minst lika bra som stepwisealgoritmen för likelihood-ratiotestet. Vi får dock vänta till modellvalideringen innan vi kan dra några slutsatser, då Ansats 1 mycket väl kan vara överanpassad till träningsdata.

Maximalt två variabler var olika mellan ansatserna. Dessa två variabler hade dessutom låg signifikans på Wald-statistikan (19), vilken skrivs ut automatiskt av SAS. Övriga variabler var antingen identiska, eller hade olika dummyvariabler från samma kategoriska variabel.

Kontroller av modellen

Bland kandidatvariablerna hade vi variablerna ”dagar mellan första föreskrivna avbetalning och observationsdatum” och ”ordning i vilken ansökningen behandlades”, för att upptäcka eventuell skevhet i data. Sådan skevhet upptäcktes ej, då variablerna ej kom med i de finala modellerna.

Vi undersökte de tre modellerna för att se om någon koefficient i β hade bytt tecken gentemot den enkla logistiska regressionen. Vi fann att alla koefficienter hade samma tecken som i den enkla logistiska regressionen.

Vi tvingade en indikatorvariabel in i stepwisealgoritmen från start. Denna variabel angav om klienten hade anställning eller ej, och kunde tänkas påverka variabler där värden saknades p.g.a. att klienten inte hade någon anställning, som anställningstid. Modellen med denna indikatorvariabel hade inte lägre medel-AR än Ansats 3.
Samma procedur upprepades för indikatorvariabeln ”medlåntagare”. Den utvecklade modellen med ”medlåntagare” hade inte heller i detta fall högre medel-AR än Ansats 3.

Vi analyserade storleken på Wald-statistikan (19) för varje variabel i varje ansats. Ingen av variablerna dominerade modellen. Det högsta värdet på Wald-statistikan var 18.6 och det minsta värdet var 3.7.

Orsaker till lågt AR

Ett AR på c:a 33% kan av erfarenhet betraktas som en något svag modell. Anledningen till detta torde vara att:

1. det endast fanns 622 dåliga kontrakt i vår datamängd. Av experter i företaget betraktas 500 dåliga kontrakt som minimum, 1000 betraktas som adekvat och 2000 som bra.
2. inga kontrakt hade mer än ett års betalningshistorik. Ett års betalningshistorik rekommenderas för att det ska visa sig vilka kontrakt som är goda och vilka som är dåliga. Vi var tvungna att gissa hur kontrakten skulle utveckla sig redan efter tre betalda avbetalningar.
3. lånet söktes ofta av klienter som inte beviljats lån av sin bank, på grund av att klienten hade betalningsproblem. Denna information fanns dock ej tillgänglig för företaget, som inte var medlem i den kreditbyrå som där de flesta banker var medlemmar i. Företaget var istället medlem i en kreditbyrå för finansföretag. Denna kreditbyrå

uppgav endast information om klienter, som hade tre avbetalningar obetalda. Det var således omöjligt att veta om klienten hade andra lån och hur han betalade amorteringarna.

Studium av residualer

Vi använde oss av tre residualplottar för att studera hur varje av de tre valda ansatserna betedde sig (jfr kapitel Residualer i Del II):

1. $\hat{h}_{ii} * \Delta\hat{\beta}_i$
2. $\hat{p}_i * \Delta\hat{\beta}_i$
3. $\hat{h}_{ii} * \Delta X_i^2$

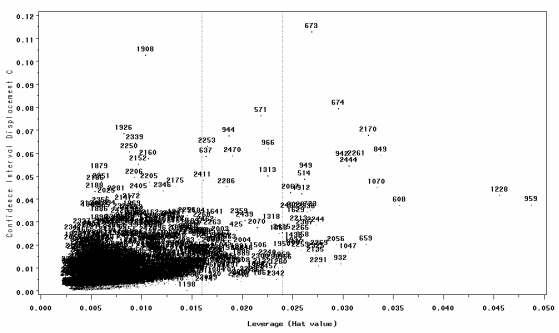
För observationer med avvikande värde i 1 till 3 ovan studerades:

4. $\Delta_l\hat{\beta}_i$, $1 \le l \le k$.

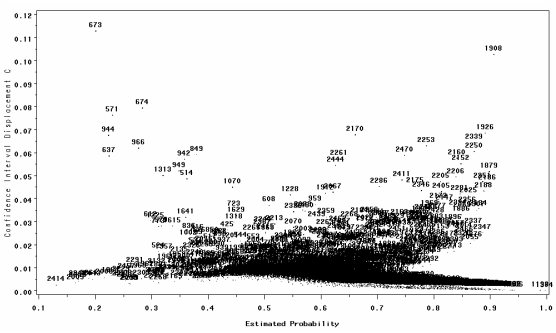
Utvecklingen av de tre Ansatserna i föregående avsnitt, föregicks av preliminära modeller, för vilka residualer studerades enligt metoden beskriven nedan. Vissa av variablerna kodades om, andra uteslöts, då de efter bedömning hade för låga p-värden eller innehöll för få observationer. När vi var nöjda med våra tre ansatser och residualerna var låga, var vi förberedda för modellvalidering.

Vi visar nedan residualplotten för den finala modellen Ansats 3, (Likelihood-ratio stepwise med WoE-kodade variabler). För alla ansatserna var residualplottarna mycket lika varandra, så vi visar endast plottarna för Ansats 3.

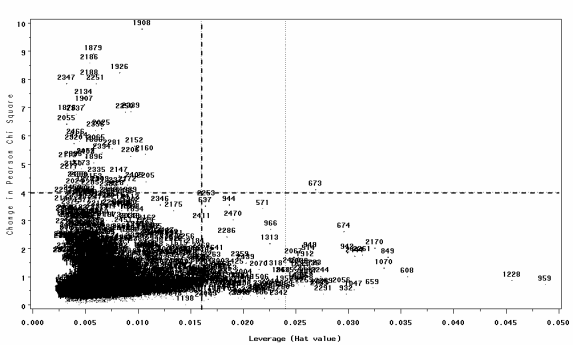
Figur 7a. $\hat{h}_{ii} * \Delta\hat{\beta}_i$



Figur 7b. $\hat{p}_i * \Delta\hat{\beta}_i$



Figur 7c. $\hat{h}_{ii} * \Delta X_i^2$



I Figur 7a ser vi att en del observationer har högt leverage ($\hat{h}_{ii}$) och relativt högt Cook-avstånd ($\Delta\hat{\beta}_i$). I Figur 7b ser vi att de flesta av dessa observationer härrör från fall där $\hat{p}_i$ redan är högt och modellen gissar rejält fel. I Figur 7c ses en vågrät streckad linje och två lodräta streckade linjer. En tumregel säger att observationer med leverage över 0.016 och ΔX_i^2 över 4 är värda att studera, motsvarande den översta vänstra delen av Figur 7c. Vi ser att endast en av observationerna ligger över båda dessa två värden.

Utifrån dessa plottar tog vi de observationer som uppfyllde villkoret:

$$\text{leverage högre än 0.16 och samtidigt } \Delta\hat{\beta}_i > 0.05 \text{ eller } \Delta X_i^2 > 3. \quad (38)$$

För dessa observationer studerade vi $\Delta_l\hat{\beta}_i,\ 1 \le l \le k$ för alla variabler i modellen. Variabler som hade $\left|\Delta_l\hat{\beta}_i\right| > 0.15$ studerades närmare och omtransformerades eventuellt. Variabler med $\left|\Delta_l\hat{\beta}_i\right| > 0.26$ omtransformerades eller uteslöts alltid.

De variabler som hade få observationer i en kategori, men gav upphov till mycket hög eller låg risk hade också högst $\left|\Delta_l\hat{\beta}_i\right|$ för de observationer som gissades fel av modellen.

Villkoret (38) uppfyllde 15 observationer för Ansats 1, 19 observationer för Ansats 2 och 19 observationer för Ansats 3. 12 av dessa observationer var gemensamma för alla tre ansatser.

Modellvalidering

Våra tre ansatser validerades på de 620 observationer, som vi lade undan innan vi började granska våra variabler (se Tabell 4). Vårt valideringsdataset är en garanti för att vi inte gjort något fel när vi utvecklade vår modell och anpassat vår modell för bra till data.

Valideringsdata är dock ej immuna mot varians i data. Vi studerade hur AR varierade i en fyrfaldig korsvalidering på modellen från Ansats 1. Genom att utföra korsvalidering på 25% av träningsdata simulerade vi på detta sätt omständigheterna under vilka valideringsdatasetet var utvaldes, där vi också valde ut 25% av den ursprungliga datamängden. Vi fick följande resultat:

Tabell 9. AR från fyrfaldig korsvalidering av Ansats 1 på träningsdata.

Korsvalidering	AR
1	38,9%
2	33,1%
3	27,3%
4	36,9%
Medel-AR	34,1%
Medelavvikelse	3.9%

Som synes från Tabell 10 får man vara försiktig när man tolkar resultatet från valideringsdata. Enskilda valideringsdata kan ge mycket olika valideringsresultat. Från Tabell 9 ser vi att valideringsdata i medeltal avviker c:a 10% från medelvärdet på AR.

Man kan inte vara säker att valideringsdata visar den riktiga styrkan hos modellen, men modellvalideringen ger dock en oumbärlig fingervisning.

Resultatet av modellvalideringen av Ansats 1, 2 och 3 visas i Tabell 10 nedan. Medel-AR visas för träningsdata och valideringsdata.

Tabell 10. Medel AR på valideringsdata och träningsdata för Ansats 1, 2 och 3.

Ansats	AR - valideringsdata	AR - träningsdata
1	30.5%	37.0%
2	34,0%	37.3%
3	33,7%	36.4%

Vi ser från ovan att Ansats 2 och 3 är har jämförbara resultat på valideringsdata. Det något lägre resultatet för Ansats 1 kan vara en indikation på att stepwisealgoritmen i SAS inte är lika stark som stepwisealgoritmen med likelihood-ratiotestet.

Då Ansats 3 innehöll färre variabler och dess variabler var något mer signifikanta än för Ansats 2, valdes Ansats 3 som final kreditvärderingsmodell.

Som jämförelse var AR 25,9%, på valideringsdata och 21.4% på träningsdata för den generiska scoremodellen, som ursprungligen användes för kreditvärderingen av vår låneportfölj (det generiska scorekortet nämns i avsnittet Beviljningsprocess ovan).

Vi studerade även AR för $p=\alpha=\delta\in\{1\%,5\%,20\%\}$ från Ansats 3 (se Figur 6) för att se om korsvalideringen, som stoppkriterium fungerade bättre än ett stoppkriterium baserat på p-värden. Resultaten visas i Tabell 11.

Tabell 11. Medel AR på valideringsdata och träningsdata för Ansats 3 med stoppkriteria $p=\alpha=\delta\in\{1\%,5\%,20\%\}$.

Stoppkriterium $p=\alpha=\delta$	AR - valideringsdata	AR - träningsdata
1%	32.9%	34.1 %
5%	33,7%	36.6%
20%	32,7%	36.4%

Vi ser från Tabell 11, att korsvalidering verkar vara ett bra stoppkriterium för stepwisealgoritmen.

Jämförelse med det generiska scorekortet

Efter en kostnadsanalys av förlusterna från ett dåligt lån och vinsterna från ett bra lån, uppskattades förlusten från ett dåligt lån vara fyra gånger större än vinsten från ett bra lån, dvs. den maximalt tillåtna risknivån sattes till 20%.

Vi ställde upp en s.k. cut-offtabell, för att hitta det $\hat{p}$ som på valideringsdata motsvarade 20% risk, se Tabell 12 nedan.

Tabell 12. Cut-offtabell: % Accepterade ansökningar, % Risk för accepterade ansökningar och vinst/ansökning på valideringsdata för Ansats 3.

					Acceptera om $\hat{p}$ < Min($\hat{p}$)		
Min($\hat{p}$)	**Max($\hat{p}$)**	**Antal kontrakt**	**Antal dåliga kontrakt**	**% Risk**	**% Accepterade ansökningar**	**% Risk för alla accepterade ansökningar**	**% Av maximal vinst per ansökning**
0.72	1	1	1	100	99.8	33.3	-66.4%
0.7	0.72	2	1	50	99.5	33.2	-65.7%
0.68	0.7	4	2	50	98.9	33.1	-64.8%
0.66	0.68	3	1	33.3	98.4	33.1	-64.5%
0.64	0.66	5	1	20	97.6	33.2	-64.4%
0.62	0.64	4	3	75	96.9	32.9	-62.5%
0.6	0.62	10	4	40	95.3	32.8	-61.0%
0.58	0.6	12	9	75	93.4	32	-56.0%
0.56	0.58	12	9	75	91.5	31	-50.3%
0.54	0.56	14	8	57.1	89.2	30.4	-46.4%
0.52	0.54	10	5	50	87.6	30	-43.8%
0.5	0.52	14	6	42.9	85.3	29.7	-41.4%
0.48	0.5	14	5	35.7	83.1	29.5	-39.5%
0.46	0.48	19	7	36.8	80	29.2	-36.8%
0.44	0.46	20	9	45	76.8	28.6	-33.0%
0.42	0.44	22	9	40.9	73.2	28	-29.3%
0.4	0.42	31	15	48.4	68.2	26.5	-22.2%
0.38	0.4	34	14	41.2	62.7	25.2	-16.3%
0.36	0.38	27	11	40.7	58.4	24	-11.7%
0.34	0.36	31	9	29	53.4	23.6	-9.6%
0.32	0.34	29	11	37.9	48.7	22.2	-5.4%
0.3	0.32	32	10	31.3	43.5	21.1	-2.4%
0.28	0.3	30	9	30	38.7	20	0.0%
0.26	0.28	32	15	46.9	33.5	15.9	6.9%
0.24	0.26	28	5	17.9	29	15.6	6.4%
0.22	0.24	32	6	18.8	23.9	14.9	6.1%
0.2	0.22	30	5	16.7	19	14.4	5.3%
0.18	0.2	31	7	22.6	14	11.5	6.0%
0.16	0.18	24	0	0	10.2	15.9	2.1%
0.14	0.16	22	3	13.6	6.6	17.1	1.0%
0.12	0.14	14	3	21.4	4.4	14.8	1.1%
0.1	0.12	13	2	15.4	2.3	14.3	0.7%
0.08	0.1	8	0	0	1	33.3	-0.7%
0.06	0.08	4	2	50	0.3	0	0.3%
0	0.06	2	0	0	0	0	0.0%

Förklaring till Tabell 12: Vi har efter att ha skattat $\hat{p}$ med modellen från Ansats 3 delat upp valideringsdata i intervall om 2% (första 2 kolumnerna). För varje intervall har vi angivit hur många kontrakt ingår och hur många av dessa kontrakt är dåliga i det aktuella intervallet (kolumn 3 och 4).

Kolumn 5: ”% Risk” ≡”Antal dåliga kontrakt” / ”Antal kontrakt” i det aktuella intervallet.

Kolumn 6. ”% Accepterade ansökningar”: Denna kolumn anger hur stor andel av låneansökningarna vi kommer att acceptera om vi sätter cut-off till min($\hat{p}$) för det aktuella intervallet, d.v.s. om vi väljer att avvisa alla ansökningar med $\hat{p}$ >min($\hat{p}$) och acceptera alla ansökningar med $\hat{p}$ <=min($\hat{p}$).

”% Accepterade ansökningar” ≡ ”Antal kontrakt med $\hat{p}$ <min($\hat{p}$)” / ”Antal kontrakt”, där min($\hat{p}$) är angivet i Kolumn 1 och ”antal kontrakt” = 620, d.v.s. alla kontrakt i valideringsdatabasen.

Kolumn 7. ”% Risk för alla accepterade ansökningar”: Denna kolumn anger hur stor risk vi kommer att ha i vår låneportfölj om vi sätter cut-off till min($\hat{p}$).

”% Risk för alla accepterade ansökningar” ≡ ”Antal dåliga kontrakt med $\hat{p}$ <= min($\hat{p}$)” / ”Antal kontrakt med $\hat{p}$ <= min($\hat{p}$)”.

Kolumn 8. ”% Av maximal vinst / ansökning”: Detta värde anger hur stor del av den maximala vinsten som erhålls om vi accepterar alla ansökningar med $\hat{p}$ <=max($\hat{p}$) och avvisar alla ansökningar med $\hat{p}$ >max($\hat{p}$). Den maximala vinsten är definierad som den vinst som skulle erhållas om alla kontrakt i validerinsdatabasen skulle accepteras och inget av dessa accepterade kontrakt skulle vara dåligt.

”% Av maximal vinst / ansökning” ≡ ”% Accepterade ansökningar” * [(100%-”% Risk för alla accepterade ansökningar”) *1 – ”% Risk för alla accepterade ansökningar” * 4)].

Hakparentesen i definitionen för ”% av maximal vinst / ansökning” anger den förväntade vinsten per kontrakt, där faktorn 4 anger att förlusten från ett dåligt kontrakt är fyra gånger större än vinsten från ett bra kontrakt.

Vi ser att den högsta förväntade vinsten för Ansats 3 är 6.9%, och att den uppnås när vi avvisar alla ansökningar med $\hat{p}$ >0.26. Vi ser att för $\hat{p}$ <=0.26 gäller mestadels att kolumn 5, ”% Risk”<20%.

Motsvarande analys utfördes på det generiska scorekortet på valideringsdata. I stället för $\hat{p}$ användes ”score-poängen” från det generiska scorekortet (se Tabell 1 för ett exempel). I övrigt var tillvägagångssättet analogt med ovan. Den högsta förväntade vinsten för det generiska scorekortet fastställdes till 3.0%.

Ansats 3 förväntas sålunda ge c:a dubbelt så stor vinst som det generiska scorekortet (6.9% mot 3.0%). Viss försiktighet är tillrådig, då vi i våra uträkningar har använt oss av uppskattade värden på relativt få data och vissa data uteslutits. Resultatet är inte desto mindre en indikation på att vinsterna förväntas öka med det nya scorekortet.

***

Appendix - Alternativa ansatser

Under modelleringsarbetet prövades ett antal alternativa metoder, vilka i korthet förtjänar att omnämnas. Resultaten från dessa modelleringsmetoder på AR var inte tillräckligt intressanta för att motivera en utförligare redovisning. Den intresserade läsaren hänvisas till litteraturen.

1. Datamatrisen **X** delades upp i två olika dataset, $\mathbf{X}^0$ och $\mathbf{X}^1$ efter värdet på en indikatorvariabel i **X** . Två logistiska regressionsmodeller utvecklades separat på $\mathbf{X}^0$ respektive $\mathbf{X}^1$. Detta förfarande upprepades för alla indikatorvariabler, se Siddiqi (2006).
2. Stepwisealgoritmen modifierades genom att inkludera svagare variabler före starkare variabler i modellen, se Siddiqi (2006).
3. Vi använde oss av alternativa algoritmer för modellering:
    a. Fractional polynomial regression med hjälp av ett SAS-makro, se Hosmer och Lemeshow (2000).
    b. Generaliserade additiva modeller baserade på kubiska splines. Vi använde paketet mgcv i R, se Wood (2006).
    c. Beslutsträd. Vi använde ett programpaket kallat för QUEST, se Loh, och Shih (1997).
4. Ett försök med samspelsvariabler utfördes. Intervallvariabler omtransformerades till intervallet [0,1]. Vi prövade därefter att inkludera de mest signifikanta samspelen. Då antalet utvärderade samspel var mycket stort, >10.000, lyckades vi inte skilja de samspel, som ökade medel-AR vid korsvalidering från brusvariabler.

Litteratur

Allison, Paul D. (1999): Logistic Regression using the SAS® system: Theory and Application, SAS Institute Inc.

Durand, David (1941): Risk Elements in Consumer Instalment Financing, National Bureau of Economic Research

Fahrmeir, Ludwig och Tutz, Gerhard (2001): Multivariate statistical modelling based on generalized linear models – 2nd ed., Springer-Verlag

Hand, David J. (2005): Good practice in retail credit scorecard assessment, The Journal of the Operational Research Society, vol. 56

Hosmer, David W. och Lemeshow, Stanley (2000): Applied Logistic Regression – 2nd ed., John Wiley & Sons, Inc.

Jabaily, Bob (Editor, 2004): Credit history: The evolution of consumer credit in America, The Ledger - The Federal Bank of Boston's Economic Education Newsletter, Spring/Summer 2004, Public and Community Affairs Department, Federal Reserve Bank of Boston

Kohavi, Ron (1995): A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2 (12)

Lilja, Kristina (2004): Marknad och hushåll – Sparande och krediter i Falun 1820-1910 utifrån ett livscykelperspektiv, Acta Universitatis Upsaliensis, Uppsala studies in Economic History, 71

Loh, Wei-Yin och Shih, Yu-Shan (1997): Split selection methods for classification trees, Statistica Sinica, vol. 7

SAS Institute Inc. (1999): SAS/STAT® User's Guide, Version 8, SAS Institute Inc.

Shtatland, Ernest S.; Cain, Emily och Barton, Mary B. (2001): The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system, Proceedings of the 26th Annual SAS Users Group International Conference

Siddiqi, Naeem (2006): Credit Risk Scorecards – Developing and Implementing Intelligent Credit Scoring, John Wiley & Sons, Inc.

Thomas, Lyn C. (2000): A Survey of Credit and Behavioural Scoring; Forecasting financial risk of lending to consumers, International Journal of Forecasting, Vol. 16, No. 2

Turney, Peter: A Theory of Cross-Validation Error (1994), Journal of Experimental and Theoretical Artificial Intelligence, 6

Vogel, David S., Gottschalk, Eric och Wang, Morgan C. (2004): Anti-matter detection: Particle Physics Model for KDD Cup 2004, SIGKDD Explorations, Volume 6, Issue 2

Wood, Simon N. (2006), Generalized Additive Models - An introduction with R, Chapman & Hall/CRC