STOCKHOLMS UNIVERSITET
Statistiska institutionen

Statistiska modeller för att avgöra förväntat utfall inom biståndsarbete

Statistical models for deciding expected results within foreign aid

Tommy Löfgren

15-högskolepoängsuppsats inom Statistik III, ht 2011
Handledare: Mikael Möller

Sammanfattning

I texten används data från biståndsmyndigheten Sida för att visa hur måluppfyllelse för insatser, kopplade till fattigdomsbekämpning, kan förklaras med hjälp av bakomliggande variabler. Totalt analyseras 358 observationer. För att bestämma sambandet appliceras multivariata metoder och olika regressionsmodeller. Slutsatsen är att en Logit-modell är bäst på att predicera utfallet (77,9 % korrekta klassificeringar). Insatslängd, intern risk, extern risk och typ av assistans har signifikant effekt på utfallet. Varken policymarkörerna, genomförandeplats eller sektor bidrar tillräckligt mycket till förklaringen för att bli inkluderade i modellen.

Nyckelord: bistånd, ordinalskala, ”Linear Discriminant Model”, ”Linear Probability Model”, ”Logit Model”, ”Probit Model”, ”Polytomous Logistic Rregression”, ”Proportional Odda Model”.

Innehåll

1: Inledning, syfte och frågeställning

Svenska biståndsbudgeten uppgick till 35,2 miljarder kronor år 2011 . Det motsvarar ca. 1 % av Sveriges bruttonationalinkomst. Biståndsmyndigheten Sida ansvarar för ungefär 30 miljarder kronor varav 946 miljoner används till förvaltning och drift (Sida 2011). Syftet med bistånd är att bidra till minskad fattigdom. Myndigheten utformar, enligt regeringens direktiv, strategier och policy för utvecklingssamarbete i olika delar av världen. Fattigdomsbekämpningen sker i form av insatser och totalt genomförs ca. 1000 stycken nya insatser per år.

Trots att biståndsarbetet omfattar svindlande summor är resultatuppföljningen begränsad. Redovisning består i huvudsak av redogörelser för hur mycket som betalats ut till olika aktörer. Sällan förekommer utvärdering gällande vilka insatser som varit lyckade respektive mindre lyckade. Den uppföljning som finns presenteras i resultatbilagan till Sidas årsredovisning där ett stratifierat slumpmässigt urval av insatser granskas i syfte att belysa särskilda teman inom biståndsarbetet (jämställdhet och kvinnors roll i utveckling var temat år 2010). Målsättningar rörande effekter och prestationer jämförs kvalitativt med uppnådda resultat. Insatserna bedöms vidare i sin helhet och tilldelas betyg på ordinalskala (som betyder att det förekommer en hierarki mellan betygsnivåerna men att avstånden inte är kvantifierbara): ”tillfredsställande”, ”mindre tillfredsställande” eller ”mycket tillfredställande”. Det ger en viss bild av vad som varit framgångsrikt men uppföljningen är begränsad till de områden som urvalen fokuserar på. Vidare analyseras inte insatserna aggregerat så att det går att se, på ett mer generellt plan, vilken typ av insatser som levererat bra resultat.

Syftet med den här uppsatsen är att identifiera en kvantitativ modell som använder (bakgrunds)information om insatserna för att skilja lyckade- från mindre lyckade insatser sett till deras måluppfyllelse. På så sätt kan aktörer inom biståndsverksamhet ta hänsyn till det förväntade utfallet för en insats givet insatsens karaktär. Naturligtvis kan en insats vara motiverad även när målen svårligen låter sig uppfyllas då potentiella förtjänster är mycket stora. Men idag saknas kvantitativa beslutsunderlag för att göra en bedömning huruvida en insats bör genomföras när hänsyn är tagen till sannolikheten att den ska gå enligt planerna. Tänk dig att en satsning på infrastruktur gör att en fattig regions tillväxt ökar med totalt 50 000 000 SEK under förutsättning att den går enligt planerna och att effekten annars är hälften så stor d.v.s. insatsens värde sjunker till 25 000 000 SEK ifall målen inte uppfylls. Anta att insatsen kostar 30 000 000 SEK. Om sannolikheten att insatsen ska gå enligt planerna är 80 % så är 30 000 000 < 50 000 000 * 80 % + 25 000 000 * 20

% = 45 000 000. Här är insatsens kostnad lägre än förväntade fördelarna. Det är ett starkt skäl till att insatsen bör genomföras. Jämförelsevis innebär en sannolikhet på 15 % att 30 000 000 > 50 000 000 * 15 % + 25 000 000 * 85 % = 28 750 000. I så fall är insatsen inte motiverad eftersom värdet av insatsresurserna är större än fördelarna. Förvisso har många insatser utfall som inte enkelt kan uttryckas i termer av pengar – vad är ett människoliv värt? – men i praktiken görs sådana avvägningar hela tiden annars skulle t.ex. vägverket placera ut vägräcken på alla vägar eftersom det sparar liv men måste prioritera vissa vägar så länge som tillgängliga medel är begränsade. Vetskap om sannolikheter gör det enklare att rationellt fördela resurserna så att biståndseffektiviteten (nytta dividerat med kostnad) blir så stor som möjligt.

Frågeställning för den här uppsatsen: Vilka förhållanden ökar sannolikheten att en biståndsinsats ska uppfylla uppsatta mål? Utifrån statistisk synpunkt är det av intresse att argumentera för den modell som är mest lämplig för att hantera det datamaterial som presenteras i nästa sektion.

Efter att datamaterialet introducerats följer ett avsnitt som lägger fram teorin bakom de modeller som används i uppsatsen. Sedan kommer ett metod-avsnitt som visar hur modellerna appliceras på data och vilka antaganden som testas. Därefter följer resultat tillsammans med analys för respektive modell och avslutningsvis sammanfattas och jämförs modellerna. I slutsatsen illustrerar även hur uppsatsens insikter kan användas i praktiken.

2: Databaserna SiRS och A+

Under hösten 2011 praktiserade jag på Sida och kom då i kontakt med en numera nerlagd databas, SiRS (Sida Rating System), som var i bruk under åren 2004 till 2007. Handläggare över hela världen betygsatte resultat för insatser samt bedömde risker. Syftet var att dokumentera, överblicka och kommunicera utfallet inom organisationen. Totalt rapporterades 358 insatser. Det motsvarar ett betydligt större antal än det urval som årsredovisningarna behandlar över samma period. Dessutom förefaller det inte finnas något särskild tendens (till exempel tematisk prioritering) bland insatser som blivit bedömda jämfört med sådana som inte betygsatts. Därför kan insatserna anses vara representativa för hela biståndssverksamheten på Sida. Det här är första gången databasen analyseras.

En aspekt att ta i beaktande är att bedömningarna gjordes något- eller några år efter att insatsen påbörjats men ännu inte avslutats. Således fanns det fortfarande utrymme att vidta åtgärder när betyget rapporterades. Därför signalerar betyget hur insatsen dittills har fortlöpt men är inte en definitiv bedömning och det är tänkbart att eventuella problem blivit lösta eller att det uppstått nya komplikationer med tiden.

Risk bedömdes dels på intern nivå i form av projektspecifika risker – relevans, effektivitet, projektdesign och hållbarhet – och dels på extern nivå utifrån insatsens makroklimat. Den generella bedömningen gjordes systematiskt enligt kriterier som handläggare tagit del av efter utbildning i systemet. Betyg rapporterades på följande fyrgradiga ordinalskala:

- Very Good (VG): ”Implementation exceeds plans in terms of quantity, quality, time or costs, without compromising the quality and/or the realization of the project purpose.”
- Good (G): “Implementation is in principal accordance with plans.”
- Acceptable (A): Implementation falls somewhat short of plans but is still on track”
- Unsatisfactory (U): “Substantial and serious shortfalls compared with plans”

Även risk bedömdes på en fyrgradig ordinalskala:

- Low (L)
- Modest (M)
- Substantial (S)
- High (H)

För att kunna bygga kvantitativa modellen har jag kopplat SiRS till en annan databas på Sida s.k. A+ som innehåller diverse information om insatserna. Jag valt ut de variabler från A+ som kan tänkas vara relevanta för att förklara samband mellan insatsens bedömning och bakomliggande förhållanden. Totalt används 13 förklaringsvariabler inklusive riskbedömningarna – två på kvotskala, sex på ordinalskala och fem på nominalskala.

2.1 Förklaringsvariabler kvotskala

- Insatslängd: antal månader insatsen är avtalad att pågå.
- Insatsstorlek: avtalat utbetalningsbelopp som insatsen omfattar.

2.2 Förklaringsvariabler ordinalskala

- Intern risk enligt ovan.
- Extern risk enligt ovan.
- ”Demokrati och mänskliga rättigheter”: graderas utifrån hur mycket insatsen syftar till att främja det här policymålet.
- ”Fred och säkerhet”: graderas utifrån hur mycket insatsen syftar till att främja det här policymålet.
- ”Jämställdhet”: graderas utifrån hur mycket insatsen syftar till att främja det här policymålet.
- ”Miljö”: graderas utifrån hur mycket insatsen syftar till att främja det här policymålet.

Policymarkörerna har värdena huvudsyfte (2), hänsyn har tagits (1) och inte relevant (0).

2.3 Förklaringsvariabler nominalskala

- Genomförandeplats: insatsens geografi på kontinentnivå.
- Förekomsten av s.k. regleringsbrev: samlingsnamn för internationella konventioner (K) och övriga direktiv (Ö) en insats kan följa t.ex. HH (hälsoinsatser inom humanitära områden), Montrealprotokollet (minskning av ozonnedbrytande ämnen), Stockholmskonventionen (utfasning av miljöföroreningar) eller A05 (HIV/AIDS-relaterade insatser).
- Genomförandekanal: aktör som genomför insatsen. Här multilaterala organisationer (FN, IMF m.fl.), svenska organisationer (universitet, privata företag, kommuner och landsting m.fl.), samarbetslandets organisationer, övriga länders organisationer eller internationella (enskilda) organisationer.
- ”Typ av assistans”: här projektstöd, programstöd, personalstöd, krediter och forskning.
- Sektor: samhällsområde (enligt OECD-DAC kodningsdirektiv) insatsen syftar till att utveckla. Här social infrastruktur, ekonomisk infrastruktur, produktionssektor, miljö/flersektorstöd och övrigt.

Datamaterialet är presenterat i sin helhet i appendix A.1.

3: Teori

För att bestämma sambandet mellan beroende utfallsvariabeln och oberoende förklaringsvariablerna finns två huvudalternativ. Det första alternativet är att förenkla beroende variabeln så att den endast kan anta två värden d.v.s. är i binär form: 1) utfall enligt planerna eller bättre (motsvarande G och VG) och 2) utfall som avviker delvis eller mycket från planerna (motsvarande A och U). Ett argument för det här tillvägagångsättet är att ”extremvärdena” VG och U förekommer relativt sparsamt i materialet (*tabell 1*). Således skulle informationsförlusten av att slå ihop variablerna vara förhållandevis begränsad.

VG	G	A	U	TOTALT
22	154	152	30	358
6,1 %	43,0 %	42,5 %	8,4 %	100 %

Tabell 1: Fördelning insatsbetyg

Fördelen med att använda en binär utfallsvariabel är att enklare modeller kan användas. Relevanta modeller är i så fall är multivariata metoden ”Linear Dicriminant Analysis” (LDA) och regressionsmodellerna ”Linear Probability Model” (LPM), ”Logit Model” (Logit) och ”Probit Model” (Probit).

Det andra alternativet är att ta hänsyn till variabelns alla fyra utfall. Om utfallens hierarki ignoreras – betygen U, A, G och VG betraktas som separata kategorier utan rangordning – så kan ”Multiple-Group Discriminant Analysis” (MLDA) samt regressionsmodellen ”Polytomous Logistic Regression” (PLR) användas. ”Proportional Odds Model” (POM), som liksom PLR är en utvecklad version av Logit, hanterar däremot ordinalskalan och tillvaratar på så sätt utfallsvariabelns fullständiga informationsinnehåll.

3.1 Linear Dicriminant Analysis & Multiple-Group Linear Discriminant Analysis

LDA och MLDA har fördelen att de är beräkningsmässigt smidigare än regressionsmodellerna och är enkla att tolka. Två eller flera grupper av observationer – som i det här fallet är betygsgrupper – åtskiljs (”diskrimineras” mellan) med hjälp av (förklarings)variabler. Grafiskt kan separationen illustreras enligt *bild 1* som utgår från två variabler och två grupper. Principen är emellertid densamma om fler variabler/dimensioner används samt ytterligare grupper.

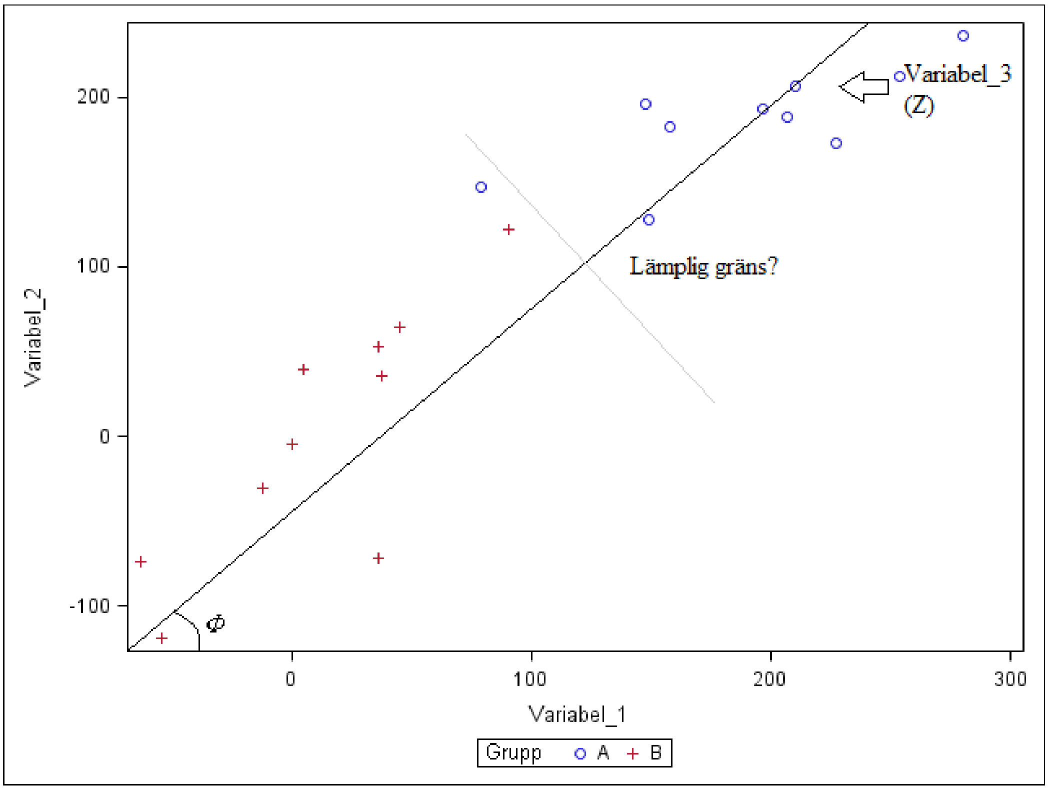

Bild1: Illustration LDA

Bilden visar att alla observationer, utom en, tillhörande grupp A har högre värde på variabel 1 jämfört med observationerna i grupp B. Värdet på variabel 2 är större för samtliga observationer i grupp A. Eftersom utfallen är uppenbart olika mellan grupperna är det motiverat att använda båda variablerna för att åtskilja grupperna. Observationerna projiceras på en tredje variabel Z, som är en linjär kombination av de andra variablerna s.k. (första) diskriminant-funktionen:

$$Z_p = w_1 * Variabel1 + w_2 * Variabel2 \quad (3.1)$$

Z_p är en punkt på nya axeln. Geometrisk logik ger att $w_1 = cos\ \Phi$ och $w_2 = sin\ \Phi$, där Φ är vinkeln mellan variabel 1 och 3. Frågan är vilken vinkel nya variabeln ska ha mot de andra variablerna och vilken gräns som är bäst på att korrekt dela in observationerna i respektive grupp. Sharma (1996: 33, 40-41, 240-242) redogör för att maximal separation mellan grupperna uppnås genom uppfyllandet av två kriterier: summan av kvadrerade avvikelser mellan grupperna ($SS_{mellan\ grupperna}$) ska vara så stor som möjlig och summan av kvadrerade avvikelser inom gruppen ($SS_{inom\ gruppen}$) ska vara så liten som möjligt. Kriterierna kombinerat innebär maximering av kvoten:

$$\lambda = \frac{SS_{\text{mellan grupperna}}}{SS_{\text{inom gruppen}}} \quad (3.2)$$

$$SS_{\text{mellan grupperna}} = \sum_{g=1}^{G} n_g (\bar{x}_g - \bar{x})^2$$

$$SS_{\text{inom gruppen}} = \sum_{g=1}^{G} [\sum_{i_g=1}^{n_g} (x_{ig} - \bar{x}_g)^2]$$

där n_g är antal observationer i grupp g, G är (totala) antalet grupper, $\bar{x}_g$ är medelvärdet för variabeln i grupp g, $\bar{x}$ är medelvärdet för variabeln för alla data, i_g är en observation i grupp g. $SS_{\text{mellan grupperna}}$ är alltså hur mycket medelvärdet för grupperna avviker från medelvärdet för hela urvalet och $SS_{\text{inom gruppen}}$ hur mycket observationerna inom grupperna avviker från respektive grupps medelvärde. Eftersom λ är en funktion av Φ gäller det att identifiera den vinkel som maximerar kvoten. Maximum kan bestämmas manuellt genom att systematiskt testa alla vinklar men statistisk programvara förenklar förfarandet. Om $\Phi = 0$ används bara variabel 1 eftersom

$$Z_p = w_1 * Variabel1 + w_2 * Variabel2 = \cos(0) * Variabel1 + \sin(0) * Variabel2$$
$$= 1\text{*Variabel1} + 0\text{*Variabel2} = Variabel1$$

och om $\Phi = 90$ används på motsvarande sätt enbart variabel 2. Vikterna i diskriminant-funktionen signalerar därför förklaringsvariablernas relativa betydelse. Dock bör tolkningen ske med försiktighet ifall variablerna är korrelerade.

Värdet på nya variabeln, för en enskild observation, kallas observationens diskriminantpoäng ("Discriminant Score"). För att avgöra vilka observationer som ska klassificeras i respektive grupp bestäms poänggränsen – insatser med lägre poäng än denna gräns placeras i ena gruppen och högre poäng ger plats i andra gruppen – så att antalet felaktiga klassifikationer minimeras. (I uppsatsen innebär felklassificering att insatser placerats i gruppen G eller VG trots att den egentligen har betyget A eller U och vise versa). Ett enkelt sätt för att bestämma adekvat gräns, när två grupper används, är att summera medelvärdet för respektive grupp $\bar{Z}_g$ och dividera summan med 2, alltså:

$$Poänggräns_{LDA} = \frac{(Z_1 + Z_2)}{2} \quad (3.3)$$

En förutsättning är att de båda grupperna är av samma storlek. Tabell 1 visar att grupperna har ungefär samma storlek (VG + G = 176 st. och A + U = 190 st.) så tillvägagångsättet är rimligt.

MLDA fungerar på samma sätt som LDA men delar in observationerna i tre eller fler grupper. För att skilja mellan grupperna kan det vara nödvändigt att identifiera ytterligare diskriminant-funktioner. Eftersom fler funktioner gör modellen mindre lätthanterlig är färre funktioner att föredra. Målet är att skilja grupperna ”tillräckligt” mycket med så få diskriminant-funktioner som möjligt. Det är tänkbart att endast en funktion ger ett tillfredsställande resultat men beroende på hur data är fördelad kan fler funktioner vara behövliga. Andra funktionen identifieras analogt genom att nya variabeln placeras i en vinkel så att λ_2 maximeras men med restriktionen att diskriminantpoängen för första och andra funktionen är okorrelerad (Sharma 1996:293). För att identifiera ytterligare funktioner *j* maximeras λ_j på motsvarande sätt givet förutsättningen att det inte förekommer korrelation med de tidigare funktionerna. Observationer tilldelas grupp beroende på poäng som beräknats för varje diskriminant-funktion.

LDA och MLDA antar att data kommer från en multivariat normalfördelaning. En vektor bestående av *j* variabler är multivariat normalfördelad om alla linjära kombinationer av vektorns dimensioner är normalfördelade, vilket i sin tur innebär att en kombination av univariat normalfördelade variabler kommer att leda till att vektorn är multivariat normalfördelade i och med att alla linjära funktioner av normalfördelade variabler automatiskt leder till att även beroende variabeln blir normalfördelad. Vidare antar modellerna att kovarians-matriserna är lika mellan grupperna. Mer om det i i metod-delen.

3.2 Linear Probability Model

LPM är snarlik en vanlig linjär (multipel) regressionsmodell (3.4) där Y_i är en binär utfallsvariabel, X_j är förklaringsvariabel *j* och β_j är regressionskoefficient *j* (*j = 1, ... k*). Totalt finns *n* (insats)observationer i datamaterialet (*i = 1, ..., n*).

$$Y_i=\beta_o+\beta_1*X_{1i}+\beta_2*X_{2i}+...+\beta_k*X_{ki}+u_i \quad (3.4)$$

3.4 kan uttryckas mer kompakt enligt:

$$Y_i = \tilde{X}_i{}' * \tilde{\beta} + u_i \quad (3.5)$$

där $\tilde{X}_i{}' = (1 \quad X_{1i} \quad \ldots \quad X_{ki})$ och $\tilde{\beta}' = (\beta_0 \quad \beta_1 \quad \ldots \quad \beta_k)$ är vektorer (notera att konstanten 1 multipliceras med interceptet).

Det förväntade värdet på Y_i är betingat av förklaringsvariablerna $\tilde{X}_i$ och $E(u_i) = 0$ d.v.s. skattningarna av regressionskoefficienterna är utan tendens (”bias”):

$$E(Y_i | \tilde{X}_i) = \tilde{X}_i{}' * \tilde{\beta} \quad (3.6)$$

Eftersom Y_i är binär följer variabeln en Bernoulli-sannolikhetsfördelning. Sannolikheten att $Y_i = 1$ är P_i och $Y_i = 0$ har sannolikheten $(1 - P_i)$. Det förväntade värdet på Y_i uttrycks i 3.7.

$$E(Y_i) = 1 * P_i + 0 * (1 - P_i) = P_i \quad (3.7)$$

En jämförelse mellan förväntade värdena i 3.6 och 3.7 ger:

$$P_i = \tilde{X}_i{}' * \tilde{\beta} \quad (3.8)$$

Således kan sannolikheten P_i att $Y_i = 1$ tolkas som en beroende variabel vilken förklaras med hjälp av variablerna i regressionsmodellen. Alltså:

$$P(Y_i = 1 | \tilde{X}_i) = \tilde{X}_i{}' * \tilde{\beta} \quad (3.9)$$

Regressionskoefficient β_j kommunicerar förändring i sannolikhet att $Y_i = 1$ när korrespondernade X_j förändras med en enhet medan alla andra förklaringsvariabler är oförändrade. För att skatta koefficienterna används minstakvadratmetoden (appendix A.2).

Det finns ett antal mer eller mindre kritiska problem med modellen (Gujarati & Porter 2009:544-547). Det mest allvarliga är att sannolikheten för en händelse naturligtvis inte kan vara mindre än noll eller större än ett, emedan regressionsmodellen inte har någon sådan begränsning. En kompromiss är att behandla $Y_i<1$ som $Y_i=0$ och $Y_i>1$ som $Y_i=1$. Sedan är feltermen u_i inte normalfördelad utan kan endast anta två värden och följer därför en Bernoulli-fördelning (jämför 3.5):

$$u_i=Y_i-\tilde{X}_i{}'*\tilde{\beta}=\begin{cases}1-\tilde{X}_i{}'*\tilde{\beta}\text{, } \mathit{när}\ Y_i=1, \mathit{med\ sannolikhet}(P_i)\\ -\tilde{X}_i{}'*\tilde{\beta}\text{, } \mathit{när}\ Y_i=0, \mathit{med\ sannolikhet}(1-P_i)\end{cases} \quad (3.10)$$

Som konsekvens går det inte att använda konfidensintervall kring regressionskoefficienterna och vidare testa ifall skattningarna är statistiskt signifikanta (punktskattningarna är användbara oavsett). Dock går det att approximera Bernoulli-fördelningen med en normalfördelning när antalet observationer är stort till följd av centrala gränsvärdessatsen. Wackerly m.fl. (2008:372) stipulerar att alla distributioner generellt kan approximeras med en normalfördelning då urvalet är större än 30 och att fördelningarna konvergerar när urvalet går mot oändligheten.

Ytterligare är problem med LPM är feltermen heteroscedastisk vilket innebär att den kvadrerade avvikelsen mellan det observerade- och skattade värdet inte är konstant d.v.s. variansen varierar. Feltermens varians härrör det faktum att termen har en Bernoulli-fördelning:

$$VAR(u_i)=P_i*(1-P_i) \quad (3.11)$$

Eftersom P_i beror av förklaringsvariablerna så är även variansen för feltermen en funktion av de här variablerna och den kommer således inte att vara konstant utan ändras systematiskt. Det leder till att skattningarna inte är effektiva eftersom de inte använder den information som det här sambandet förmedlar. En effektiv skattning minimerar variansen genom att anpassa modellen så bra som möjligt vilket alltså inte är fallet så länge som modellen kan förbättras genom att ta hänsyn till sambandet. Som tur är går det att förhållandevis enkelt att hantera problemet med heteroscedastisitet genom att transformera modellen. Mer om det i metod-delen.

Det är befogat att diskutera huruvida sannolikheten verkligen ökar linjärt. Har en månads ökning i avtalslängd samma effekt på sannolikheten när förklaringsvariablerna (kollektivt) redan har givet en hög skattning av sannolikheten, som när sannolikheten ligger nära 50 %? Jag återkommer till det nedan.

3.3 Logit model

Logit använder en s.k. logistisk funktion – istället för en linjär funktion som i LPM – för att uttrycka hur sannolikheten beror av förklaringsvariablerna (skattningarna antas åter vara utan tendens d.v.s. $E(u_i)=0$):

$$P(Y_i=1|\tilde{X}_i)=\frac{1}{[1+\exp(-\tilde{X}_i'*\tilde{\beta})]}=\frac{[\exp(\tilde{X}_i'*\tilde{\beta})]}{[1+\exp(\tilde{X}_i'*\tilde{\beta})]} \quad (3.12)$$

Den här modellen har fördelen att utfallet på beroende variabeln bara kan variera mellan noll och ett eftersom $P\rightarrow 1$ när $\tilde{X}'*\tilde{\beta}\rightarrow\infty$ och $P\rightarrow 0$ när $\tilde{X}'*\tilde{\beta}\rightarrow-\infty$. En annan aspekt är att P förändras mindre till följd av förändringar i förklaringsvariablerna när sannolikheten för ett positivt utfall ”redan är” är mycket stor respektive när den är mycket liten. Praktiskt betyder det att P förblir nära ett respektive noll när $\tilde{X}'*\tilde{\beta}$ är tillräckligt stor eller liten. Det förefaller vara troligt att förändringar i variablerna har större effekt på sannolikheten om en insats väger mellan att klassificeras som enligt planerna eller bättre, än om den redan har en hög sannolikhet att bedömas på ett visst sätt. Således kan (men är inte nödvändigtvis) en logistisk funktion mer lämplig, jämfört med en linjär sådan, för att bestämma förändring i sannolikhet.

En mekanisk tolkning av regressionskoefficierna är inte speciellt tilltalande. Tolkningen blir ännu mer komplex till följd av det faktum att förändring i sannolikhet, när en förklaringsvariabel ändras, skiljer sig beroende på vilken sannolikhet som redan är för handen i och med att förändringshastigheten inte är linjär. Sambandet kan uttryckas med en deriveringskvot i vänsterledet (Gujarati & Porter 2009:555):

$$\frac{dP_i}{dX_j}=\beta_j*P_i*(1-P_i) \quad (3.13)$$

En mer intuitiv tolkning av regressionskoefficienterna ges om en oddskvot (3.14) används istället för sannolikhet som beroende variabel.

$$\text{Oddskvot} = \frac{P_i}{(1-P_i)}=\frac{[\exp(\tilde{X}_i'*\tilde{\beta})]}{[1+\exp(\tilde{X}_i'*\tilde{\beta})]}/\frac{1}{[1+\exp(\tilde{X}_i'*\tilde{\beta})]}=\exp(\tilde{X}_i'*\tilde{\beta}) \quad (3.14)$$

Oddskvoten är sannolikheten att $Y_i=1$ relativt $Y_i=0$. Om den är större än ett betyder det att sannolikheten att händelsen ska inträffa (här: att en insats ska bedömas som enligt planerna eller bättre) är större än att den inte ska inträffa. Ifall den däremot är mindre än ett gäller det motsatta.

Anta att $P_i=0{,}7$, då är oddskvoten $\frac{0{,}7}{(1-0{,}7)}=2{,}33$ d.v.s. det är 2,33 gånger så troligt att händelsen inträffar som att den inte gör det. Om t.ex. förklaringsvariabel X_1 ökar med en enhet medan övriga förklaringsvariabler är oförändrade så påverkar det oddskvoten på följande sätt:

$$\exp[\beta_1*(X_{1i}+1)+\beta_2*X_{2i}+...+\beta_k*kX_{ki}]=\exp[\beta_1*(X_{1i}+1)]*\exp[\beta_2*X_{2i}]*...*\exp[\beta_k*kX_{ki}]$$

$$=\exp[\beta_1]*\exp[(\beta_1*X_{1i})]*\exp[\beta_2*X_{2i}]*...*\exp[\beta_k*kX_{ki}]=\exp[\beta_1]*\exp[\tilde{X}_i{}'*\tilde{\beta}]$$

Oddskvoten ökar alltså med $\exp[\beta_1]$ och det gäller oavsett nivån på förklaringsvariablerna. Således är förändringseffekten lika stor hela tiden, att jämföra med fallet där sannolikhet används som beroende variabel, vilket gör tolkningen av koefficienten entydig och smidigt kommunicerbar.

För att skatta parametrarna används (obetingad) maximum likelihood-skattningsmetod (MLE) för data på individuell nivå. Minstakvadratmetoden är inte tillämpbar eftersom modellen är icke-linjär i parametrarna. Appendix A.3 presenterar tillvägagångsättet. Stock & Watson (2007:393) redogör för att MLE ger skattningar som är utan bias, konsistenta – vilket betyder att skattningen av parametrarna konvergerar mot deras faktiska värde – och effektiva då ingen annan metod presterar skattningar med lägre varians d.v.s. är mer precisa i det här fallet. Feltermerna kring regressionskoefficienterna kan antas vara normalfördelade då antalet observationer är stort.

3.4 Probit model

Sannolikheten att $Y_i=1$ för Probit är baserad på kumulativa fördelningsfunktionen för en standardiserad normalfördelning Φ ($E(u_i)=0$):

$$P(Y_i=1|\tilde{X}_i)=\Phi(\tilde{X}_i{}'*\tilde{\beta})=\Phi(z) \quad (3.15)$$

Efter att värdet på z beräknats som en linjär funktion av förklaringsvariablerna kan sannolikheten identifieras med hjälp av en tabell över Φ som anger sannolikheten att $Z\leq z$ d.v.s.

$\Phi(z)=P(Z\leq z)$ (t.ex. Stock & Watson 2007:755). Notera att z kan vara både positiv och negativ. Eftersom det rör sig om en standardiserad normalfördelning kan sannolikheten enbart variera mellan noll och ett. $P\rightarrow 1$ när $\tilde{X}'*\tilde{\beta}=z\rightarrow\infty$ och $P\rightarrow 0$ när $\tilde{X}'*\tilde{\beta}=z\rightarrow-\infty$. Dessutom ökar sannolikheten som mest kring 50 % (då $z=0$) vilket, som nämnt, förefaller vara mer realistisk än att den ökar linjärt på det sätt LPM implicerar.

Regressionskoefficienterna är svåra att tolka på ett meningsfullt sätt och effekten av en förändring i en förklaringsvariabel, medan övriga variabler är oförändrade, undersöks bäst genom att sannolikheten före förändringen jämförs med sannolikheten efter förändringen (Stock & Watson 2007:392). Precis som i fallet med Logit varierar en förklaringsvariabels påverkan beroende på variablernas (kollektiva) nivå – förändringen, när z är stor eller liten, är mindre än när z är kring noll.

Eftersom modellen är icke-linjär är inte minstakvadratmetoden tillämpbar. MLE fungerar dock och har samma attraktiva egenskaper som vid skattningen av Logit. Tillvägagångssättet för skattningen redogörs för i appendix A.4. Parameterskattningarnas signifikans kan testas under antagandet att feltermen är normalfördelad.

3.5 Polytomous Logistic Regression & Proportional Odds Model

PLR bestämmer hur förklaringsvaribalerna relaterar till en icke-numerisk beroende variabler med tre- eller fler svarskategorier. I den här uppsatsen motsvarar svarsvaribeln fyra insatsbetygen: U, A, G, VG. Som tidigare nämnt ignorerar PLR den naturliga hierarki som betygen kan rangordnas enligt.

PLR analyserar flera logit funktioner aggregerat (Kleinbaum m.fl 2008:636-645):

$$\frac{[P(Y_i=g|\tilde{X}_i)]}{[P(Y_i=0|\tilde{X}_i)]}=\exp(\tilde{X}_i'*\tilde{\beta}_g) \quad (3.16)$$

där *g* anger vilken logit funktion det rör sig om – antalet logit-funktioner är lika många som beroende variabelns kategorier exklusive referenskategorin (så här är *g = 1, 2, 3*) – och $\tilde{\beta}_g$ är regressionskoefficenterna för logit funktion *g*. Vänsterledet uttrycker en kvot som är ”snarlik” oddskvoten och ska tolkas som sannolikheten att $Y_i = g$ relativt sannolikheten att $Y_i = 0$, där 0 motsvarar referenskategorin. När betyget U används som referenskategori blir modellen:

$$\begin{cases} \dfrac{[P(Y_i=VG|\tilde{X}_i)]}{[P(Y_i=U|\tilde{X}_i)]}=\exp(\tilde{X}_i{}'*\tilde{\beta}_1) \\ \dfrac{[P(Y_i=G|\tilde{X}_i)]}{[P(Y_i=U|\tilde{X}_i)]}=\exp(\tilde{X}_i{}'*\tilde{\beta}_2) \\ \dfrac{[P(Y_i=A|\tilde{X}_i)]}{[P(Y_i=U|\tilde{X}_i)]}=\exp(\tilde{X}_i{}'*\tilde{\beta}_3) \end{cases}$$

Notera att parametrarna i vänsterledet skiljer sig när olika kategorier jämförs med referenskategorin och att all data analyseras aggregerat i en enhetlig modellstruktur d.v.s. man skattar inte modellen som tre oberoende modeller med binära utfallsvariabler. Polynoma modellen kan ha andra skattningar än när separata modeller används och leder generellt till att ”styrkan” (Power) hos statistiska test ökar (Kleinbaum m.fl 2008:636) – alltså chansen att förkasta nollhypotsen, när den faktisk är falsk, blir större.

Ett exempel: anta att förklaringsvariabel X_2 är insatslängd, som ökar med en enhet, alltså med en månad (se avsnitt 3.1.). Om $\exp(\beta_{12})=0{,}5$, $\exp(\beta_{22})=0{,}75$ och $\exp(\beta_{32})=1$ där första ”nersänkta” siffran hos koefficienten uttrycker logit funktion *g* och andra siffran vilken variabel *j* det rör sig om (här variabel 2) så betyder det att oddset att insatsen ska bedömas som VG, G respektive A relativt U – givet att allting annat är lika – uppskattas minska till hälften så troligt för VG; minska till tre fjärdedelar så troligt för G; samt är oförändrat för A. Precis som i fallet med logit är förändringen i odds oberoende av den sannolikhetsnivå som förklaringsvariablerna implicerat före förändringen.

För POM (Kleinbaum m.fl 2008:646-650; O'Connel 2006:27-41) relateras sannolikheten att ett betyg ska inträffa till sannolikheten att något av de andra ”undre” betygen ska inträffa. Det betyder att t.ex. sannolikheten att insatsen ska bedömas som VG divideras med sannolikheten att insatsen får något av betygen G, A eller U. Om det är av intresse att veta sannolikheten att G ska inträffa, så undersöks sannolikheten att antingen G eller VG inträffar i relation till sannolikheten att antingen A eller U inträffar. Eftersom kategorierna antas följa en rangordning är det logiskt att sannolikheten för G även inkluderar sannolikheten för VG eftersom det är en förutsättning att ”åtminstone” G sker för att VG ska bli aktuell. Annorlunda uttryckt kan A och G betraktas som mellannivåer före VG som uppnås successivt när sannolikhet ackumuleras enligt 3.17.

$$P(Y_i \geq g | \tilde{X}_i) = \frac{1}{1+\exp[-(\beta_g + \beta_1 * X_{1i} + \beta_2 * X_{2i} + ... + \beta_k * X_{ki})]} \quad (3.17)$$

Modellen är en version av logit (jämför 3.12) men med skillnaden att likhetstecknet är ersatt med ett olikhetstecken och interceptet β_g varierar bereonde på den ”ordinala” kategori *g* som *Y* ska vara större eller lika med (istället för antal logit funktioner är *g* nivåer från referenskategorin vilken är ”längst ner i hiearkin” och även här är *g = 1, 2, 3* eftersom maximala antalet nivåer är tre, nämligen mellan U och VG: $U \rightarrow (1) A \rightarrow (2) G \rightarrow (3) VG$). (Vektorer används inte i högerledet så att interceptet ska bli synligt och $E(u_i)=0$).

I jämförelse med PLR så varierar inte koefficienterna, bortsätt från interceptet, när olika kategorier jämförs. Det betyder att regressionskoefficient β_j , som korresponderar med förklaringsvariabel X_j , hela tiden är samma i modellen. Effekten på sannolikheten till följd av en förändring i en förklaringsvariabel är således lika stor oavsett om vi undersöker sannolikheten att $Y_i \geq U$ som när vi undersöker sannolikheten att $Y_i \geq VG$. Antagandet att koeffecienterna, förutom interceptet, är samma kallas för ”proportional odds assumption” (därav namnet på modellen). Antagandet är kritiskt för modellens giltighet och behöver kontrolleras. I metod-delen beskrivs tillvägagångsättet.

För att tolka koefficienterna är, liksom för logit, oddskvoten användbar:

$$\text{Oddskvot}_g = \frac{[P(Y_i \geq g | \tilde{X}_i)]}{[1-P(Y_i \geq g | \tilde{X}_i)]} = \exp(\beta_g + \beta_1 * X_{1i} + \beta_2 * X_{2i} + ... + \beta_k * X_{ki}) \quad (3.19)$$

För att skatta parametrarna används MLE på anlogt sätt som vid skattningen av Logit som beskrivs i appendix.

4: Metod och modeller

Statistiska programvaran SAS 9.2 används för att genomföra beräkningar och parameterskattningar på det sätt som beskrivits i teori-avsnittet och i appendix.

För att avgöra vilka förklaringsvariabler som ska ingå i respektive modell är det till att börja med nödvändigt att ta hänsyn till att multivariata modellerna LDA och MLDA ställer kravet att förklaringsvariablerna tillsammans är multivariat normalfördelade och att grupperna har lika kovarians-matriser. Sharma (1996: 263) kommenterar att antagandet om multivariat normalfördelning måste vara uppfyllt för att det ska gå att testa diskriminant-funktionens statistiska signifikans och att klassificeringsresultaten annars riskerar att vara missvisande. Om kovarians-matriserna är olika ger det likvärdiga effekter på resultatet. Dock påpekar han att modellerna är förhållandevis robusta. Kvalitativa variabler är oförenliga med antagandet om normalfördelning eftersom multivariat normalfördelning i regel förutsätter att respektive variabel är univariat normalfördelad. Därför tycks de fem variblerna på nominalskala (se sektion 2) inte speciellt användbara . Johnson & Wichern (2007:644) hänvisar dock till datasimuleringar av diskriminant-modeller som visat att en kombination av kvantiativa och kvalitativa variabler kunnat leverera tillfredsställande resultat. Men jag har valt att inte inkludera variabler på nominalskala i LDA och MLDA. Angående sex varibalerna på ordinalskala (se sektion 2) så går det inte att anta en normalfördelning så länge som avstånden mellan nivåerna inte är kvantifierbara. Därför har även de fått utgå. Två variabalerna på kvotskala är däremot fullgoda kandidater för modellerna. Jag kommer att testa antagandet om multivariat normalfördelning på ett sätt som är snarlikt grafiska granskningen av univariat normalfördelning då s.k. ”Q-Q Plot” används. Sharma (1996: 380-382) redogör för tillvägagångsättet: först beräknas respektive observations Mahalanobis avstånd till observationernas geometriska centrum i kvadrat och avstånden följer en chi-två-fördelning när $n>25$; avstånden sorteras därefter från lägt till störst; sedan beräknas den percentil som varje observation *i* tillhör enligt $(i-0{,}5)/n$; chi-två-värdet för respektive percentil hämtas avslutningsvis från en vanlig chi-två-tabell och jämförs grafiskt med sorterade avstånden. Om sambandet är linjärt, så att avstånden är fördelade på samma sätt som en teoretisk normalfördelning, är antagandet om multivariat normalfördelning uppfyllt. Antagandet att kovarians-matriserna är lika kommer att granskas och undersökas formellt genom ett chi-två test.

Diskriminant funktione/funktionerna kommer beräknas med hänsyn tagen till gruppernas relativa storlek. Det betyder att andelen observationer som placeras i en grupp är proportionerlig mot antalet observationer i datamaterialet som tillhör den gruppen. Följande diskriminant-funktion(er) ska skattas:

$$Z_p = w_1 * Insatslängd + w_2 * Insatsstorlek$$

Ju fler korrekta klassifikationer desto bättre modell. För att avgöra vilken modell som är bäst på att hantera datamaterialet i uppsatsen kommer jag att jämföra LDA och MLDA med regressionsmodellernas kapacitet (förutom PLR) att göra korrekta prediktioner.

Regressionsmodellerna LPM, Logit, Probit, PLR och POM ställer till skillnad från LDA och MLDA inga särskilda krav på förklaringsvariablernas fördelning. Men eftersom antalet variabler är så pass stort är det nödvändigt att rensa bort sådana variabler som inte bidrar mycket till förklaringen av variationen hos beroende variabeln. Generellt är en modell med färre variabler att föredra eftersom överflödiga förklaringsvariabler, som varken är av statistisk eller ”praktisk” signifikanta, bara förvirrar användaren av modellen och dessutom leder till ökad risk att s.k. ”multicollinearity” uppstår bland varibalerna d.v.s. att det finns ett starkt linjärt samband mellan två eller flera av dem. En s.k. ”stepwise”-procedur gör att onödiga variabler bortprioriteras från modellen. Först inkluderas den variabel som ger modellen störst $F - v\ddot{a}rde = \frac{\text{Variation i Y som modellen förklarar}}{\text{Variation i Y som modellen inte förklarar}}$, varav F är en F-fördelad test-statistiska där nollhypotsen är att förklaringsvariablerna inte förklarar ett statistiskt signifikant andel av variationen i Y. Ett lågt p-värde innebär att nollhypotsen kan avfärdas med stor säkerhet eftersom p-värdet anger sannolikheten att ha ett så extremt F-värdet trots att nollhypotsen är korrekt. Ju lägre p-värde desto mer orimligt är det att nollhypotsen är sann och att observerade F-värdet beror på slumpen. Efter att första variabeln introducerats undersöks alla återstående förklaringsvärdes partiella F-värde, som är den förändring i F-värdet som introduktion av dem skulle förorsaka givet alla variblerna som redan befinner sig i modellen. Den variabel som ökar F-värdet mest (vilket innebär att den har lägst p-värde) introduceras förutsatt att p-värdet är ”tillräckligt” lågt. Det är upp till en själv att besluta vilken storlek på p-värdet som ska användas som gräns, alltså den signifikansnivå som krävs för att varibeln ska få ingå i modellen. Konventionella nivåer är 1 %, 5% eller 10 %. Jag har dock valt att vara lite generös och använder 15 % som nivå d.v.s. jag tillåter att 15 % av variationen som förklaras av variabeln i genomsnitt beror på slumpen. Förutom stegvis inkludering av nya variabler så exkluderar stepwise-proceduren, efter varje steg, sådana variabler som ingår i modellen men som inte längre bidrar signifikant till förklaringen av variationen hos beroende variabeln sedan nya variabler introducerats. En anledning till att variabelns betydelse varierar vid olika stadier kan vara att det förekommer multicollinearity så att nya variabler i modellen kan vara starkt korrelerade med tidigare variabler och därför gör dem mindre betydelsefulla i och med att de kan ersätta varandra. Proceduren avslutas när inga signifikanta variabler kan inkluderas eller icke signifikanta sådana exkluderas från modellen.

För att använda variabler på nominalskala i LPM är det nödvändigt att introducera en uppsjö med dummy-varibler, alltså sådana som bara antar värdena ett och noll. Variabeln regleringsbrev har t.ex. tre kategorier: K, Ö och regleringsbrev saknas (se sektion 2). Om insatsen har ett regleringsbrev K så kodas variabeln som ett annars noll och samma princip gäller för regleringsbrev Ö. Dock ingår inte tredje kategorin, regleringsbrev saknas, i modellen eftersom det skulle leda till ”dummy-variabel-fällan” (förutsatt att det finns ett intercept i modellen) som innebär att det uppstår perfekt multicollinearity mellan variablerna då det finns ett exakt linjärt samband mellan dem. När både K och Ö har värdena noll är det samma sak som att den tredje kategorin är för handen (som ingår i modellens intercept) och därför är det inkorrekt att ha med den som en oberoende separat variabel. Således kommer antalet dummy-variabler alltid att vara en färre än antalet kategorier varav den ”uteslutna” dummy-variabeln fungerar som referenskategori och effekten av de andra är skillanden i förhållande till den. Variablerna på ordinalskala ligger åter i gränszonen mellan att vara kvalitativa och kvantitativa . Det är tänkbart att de uppträder som om det fanns en intervallskala men eftersom antalet nivåer bara är fyra eller färre, har jag valt att modellerna dem som dummy-variabler d.v.s. som oberoende kategorier. Nackdelen med att många dummy-variabler är att de konsumerar lika många frihetsgrader. Frihetsgrader är antalet oberoende observatioer som kan användas för att skatta parametrarna och följande villkor måste vara uppfyllt $n-k-1>0$ där n är antalet observationer, k antalet variabler i modellen och ettan motsvarar interceptet . I det här fallet då $n \gg k+1$ är många dummy variabler inget större problem.

Som tidigare nämnt kommer sannolikhet över ett att betraktas som ett och ”negativ” sannolikhet att ses som noll i LPM. Gujarati & Porter (2009:545) beskriver hur problemet med homoscedastisitet kan hanteras genom att variablerna transformeras på följande sätt:

$$\frac{Y_i}{(w_i)^{(1/2)}} = \frac{(\tilde{X}_i' * \tilde{\beta} + u_i)}{(w_i)^{(1/2)}} \quad (3.20)$$

där w_i kan skattas med:

$$\hat{w}_i = \hat{Y}_i * (1 - \hat{Y}_i)$$

Det betyder att beroende variabeln (som alltså är sannolikheten att insatsen ska bedömas som enligt planerna eller bättre) till att börja med skattas trots problemet med heteroscedastisitet och därefter användas för att ta fram den transformerade modellen av LPM som i sin tur skattas.

När $P_i \geq 50$ är modellen bestämd att predicera att $Y_i = 1$ annars att $Y_i = 0$. För att utröna hur väl modellen fungerar jämförs förväntade utfallet med det faktiska utfallet. Ju fler korrekta förutsägelser desto bättre modell. Samtidigt är det viktigt att ta hänsyn till att om en insats bedöms ha sannolikheten 51 % kommer den att bedömas likadant som en insats med sannolikhet 99 %. Därför är det motiverat att handläggare använder förväntade sannolikheten som beslutsunderlag för att avgöra huruvida en insats ska genomföras, snarare än enbart förlitar sig på att utfallsvariabeln bedöms vara ett eller noll. På så sätt fås en mer nyanserad bild.

Följande modell skattas:

$$\begin{gathered} P(Y_i=1)=\beta_0+\beta_1*X_1+\beta_2*X_2+\beta_3*X_3+\beta_4*X_4+\beta_5*X_5+\beta_6*X_6+\beta_7*X_7+\beta_8*X_8+\beta_9*X_9 \\ +\beta_{10}*X_{10}+\beta_{11}*X_{11}+\beta_{12}*X_{12}+\beta_{13}*X_{13}+\beta_{14}*X_{14}+\beta_{15}X_{15}+\beta_{16}*X_{16}+\beta_{17}*X_{17}+\beta_{18}*X_{18} \\ +\beta_{19}*X_{19}+\beta_{20}*X_{20}+\beta_{21}*X_{21}+\beta_{22}*X_{22}+\beta_{23}*X_{23}\beta_{24}*X_{24}+\beta_{25}*X_{25}+\beta_{26}*X_{26}+\beta_{27}*X_{27} \\ +\beta_{28}*X_{28}+\beta_{29}*X_{29}+\beta_{30}*X_{30}+\beta_{31}*X_{31}+\beta_{32}*X_{32}+\beta_{33}*X_{33}+u_i \end{gathered}$$

där:

$$\begin{gathered} X_1=insatslängd,\ X_2=insatsstorlek,\ X_3=inter\ risk_M,\ X_4=inter\ risk_H,\ X_5=inter\ risk_S, \\ X_6=extern\ risk_M,\ X_7=extern\ risk_H,\ X_8=extern\ risk_S, \\ X_9=policy\ demokrati\ och\ mänskliga\ rättigheter_1,\ X_{10}=policy\ demokrati\ och\ mänskliga\ rättigheter_2, \\ X_{11}=policy\ fred\ och\ säkerhet_1,\ X_{12}=policy\ fred\ och\ säkerhet_2,\ X_{13}=policy\ jämställdhet_1, \\ X_{14}=policy\ jämställdhet_2,\ X_{15}=policy\ miljö_1,\ X_{16}=policy\ miljö_2,\ X_{17}=genomförandeplats_{global}, \\ X_{18}=genomförandeplats_{\text{Asien\&Östeuropa}},\ X_{19}=genomförandeplats_{\text{Syd/Central Amerika}},\ X_{20}=regleringsbrev_K, \\ X_{21}=regleringsbrev_Ö,\ X_{22}=genomförandekanal_{\text{multilaterala org.}},\ X_{23}=genomförandekanal_{\text{övriga länders org.}} \\ ,X_{24}=genomförandekanal_{\text{samarbetslandets org.}},\ X_{25}=genomförandekanal_{\text{svenska org.}},\ X_{26}=typ\ av\ assistans_{krediter}, \\ X_{27}=typ\ av\ assistans_{personalstöd},\ X_{28}=typ\ av\ assistans_{programstöd},\ X_{29}=typ\ av\ assistans_{projektstöd}, \\ X_{30}=sektor_{\text{miljö/flera}},\ X_{31}=sektor_{övrigt},\ X_{32}=sektor_{produktion},\ X_{33}=sektor_{\text{social infrastruktur}} \end{gathered}$$

För intern- och extern risk används L (låg risk) som referenskategori, samtliga policymarkörer har värdet noll (inte relevant) som referenskategori, genomförandeplats har Afrika som referenskategori, regleringsbrev har regleringsbrev saknas som referenskategri, genomförandekanal har internationella organisationer som referenskategori, typ av assistans har forskning som referenskategori och sektor har ekonomisk infrastruktur som referenskategori.

Logit, probit, PLR och POM använder samma förklaringsvariablerna som LPM men sannolihetsfunktionen skiljer sig på det sätt som beskrivits i teori-avsnittet. Stepwise-proceduren används genomgående för att identifiera vilka variabler som bidrar ”mycket” till att förklara variationen i beroende variabeln. Förutom parameterskattningen presenteras även regressionskoefficnternas effekt på oddskvoten för Logit.

Logit och Probit bedömmer $Y_i = 1$ när sannolikheten är över 50 % och annars är $Y_i = 0$. Eftersom förändring i sannolikhet är som störst kring 50 % kommer antagligen färre observationer att ligga precis på gränsen och därför kan handläggare i större utsträckning, än vid tillämpandet av linjära modellen, följa utfallsvariabeln för att avgöra ifall en insats kan förväntas gå enligt planerna eller bättre. För att avgöra hur väl modellerna fungerar är det åter rimligt att undersöka hur många prediktioner som är korrekta relativt antalet felaktiga sådana. POM har också en sådan kvot och predicerade betyget beräknas som en sammanvägning av sannolikheter att respektive ordinala kategori ska inträffa (SAS 2011). Det bör poängteras att det inte går att bedöma hur väl PLR fungerar utifrån måttet andel korrekta klassifikationer eftersom det rör sig om separata funktioner. Oavsett är det av intresse att se hur olika värden på förklaringsvariablerna uppskattas påverka sannolikheten att en insats ska bedömas som t.ex. A relativt referenskategorins betyg U.

POM förutsätter att regressionskoefficinetera är konstanta oavsett ordinal kategori s.k. ”score test” undersöker huruvida villkoret är uppfyllt. ”Test-statistikan” är chi-två-fördelad med $k(g-2)$ frihetsgrader, där k är antalet förklaringsvariabler i modellen. Nollhypotesen är att parametrarna (förutom interceptet) är lika. Om hypotesen förkastas betyder det att modellen inte är giltig och i så fall bör PLR användas istället för POM.

En kort kommentar kring oberoende: det är tänkbart att en del observationer inte är helt oberoende av varandra till följd av att handläggare administrerar flera insatser. Totalt förekommer 137 olika handläggare varav 6 har betygsatt fler än 10 insatser men ingen har betygsatt fler än 20. Det finns ingen statistisk teknik för att hantera oberoende men resultatet kan möjligen påverkas av att en överrepresenterad handläggare har en tendens att bedöma på ett visst sätt och att handläggare som inte valt att rapportera betyg har andra egenskaper en sådana som var aktiva med Sirs.

Avslutningsvis bör det påpekas att regressionsmodeller generellt är förknippade med några grundläggande antaganden: förutom avsaknaden på multicollinearity och heteroskedasticitet, som vi varit inne på, så riskerar resultaten att påverkas om det förekommer autokorrelation i datamaterialet d.v.s. feltermerna är korrelerad med tidigare feltermer i en serie. Men autokorrelation är synnerhet ett problem som rör data av tidsserie-typ och därför berör jag inte autokorrelation vidare. Mer specifikt kring multicollinearity så leder sådan till skattningarna blir mindre precisa genom att variabeln ökar vilket även påverkar sannolikheten att nollhypotsen – att varibeln inte har någon signifikant påverkan – i större utsträckning kommer att behållas. Det leder till att stepwise-proceduren kommer att ta bort fler variabler än ifall variablerna inte hade något samband. Genom att använda ett mått som kallas VIF (”variance-inflating factor”) går det studera i vilken utsträckning multicollinearity existerar bland variablerna. Antigen studeras hur två varibaler påverkar varandra t.ex. variabel två och tre:

$$VIF = \frac{1}{\left(1 - r_{23}^2\right)} \quad (3.21)$$

eller så används R_j^2 , som anger är hur mycket av variationen i förklaringsvariabel *j* som kan förklaras av övriga variabler i modellen, alltså linjära sambandet mellan den och övriga variabler:

$$VIF = \frac{1}{\left(1 - R_j^2\right)} \quad (3.22)$$

När $R_j^2 \to 1$, så att all variation kan förklaras, så betyder det att $VIF \to \infty$. Ifall $R_j^2 \to 0$ så leder det till att $VIF \to 1$. Storleken på VIF är direkt proportionerlig mot variansen hos skattade koefficienten β_j , vilket gör konfiendensintervallet större när R_j^2 ökar och således blir förklaringsvariabeln mindre användbar. 3.22 kommer att användas för att avgöra graden av multicollinearity för respektive variabel gentemot övriga varibaler kollektivt. Jag misstänker att VIF för alla variabler på nominal- och ordinalskala (alla behandlas som kategoriska variabler) kommer att vara av begränsad storlek när det tas i beaktande att R^2 enligt Gujarati & Porter (2009:546) i regel blir låg – ofta mellan 0,2 och 0,6 – då en binär variabel används som beroende variabel.

Angående heteroskedasticitet så har vi redan diskuterat hur LPM kan korrigeras för att komma till bukt med det problem som modellens struktur föranlett. I övrigt är det ganska svårt att åtgärda den icke konstanta varians som kan förekomma. Whites heteroskedasticisk-konsistenta ”robusta standardfel” tillhandahåller konstanta standaravvikelser (kvadratroten ur variansen) när antalet observationer är stort, som i det här fallet. Proceduren för att beräkna robusta standarfel är förhållandevis komplicerad, begränsad till minstakvadratmetoden och baseras på en approximation av variansen för varje observation med kvadrerade feltermen (Gujarati & Porter 2009:391, 411). Dock är metoden väl etablerad bland praktiker och jag kommer att presentera robusta standarfel – som ofta är större än ursprungliga standardavvikelsen hos skattningen – för att inte dra felaktiga slutsatser orsakad av potentiell heteroskedasticitet.

5: Resultat och analys

5.1 Multicollinearity

Parameter Estimates

Variable	DF	Variance Inflation
Intercept	1	0
Insatslangd	1	1.54036
Insatsstorlek	1	1.23171
internM	1	1.61808
internH	1	1.58667
internS	1	1.50878
externM	1	1.66785
externH	1	1.69605
externS	1	1.61368
demokrati1	1	2.65303
demokrati2	1	2.96010
fred1	1	1.08861
fred2	1	1.07654
jamstalldhet1	1	1.79703
jamstalldhet2	1	1.54266
miljo1	1	1.49857
miljo2	1	1.72877
Global	1	1.39668
AsienEuropa	1	2.48508
SCAmerika	1	1.85765
RegleringsbrevK	1	1.30786
Regleringsbrevo	1	1.21816
GMulti	1	3.13450
GOvriga	1	2.07201
GSamar	1	3.87911
GSven	1	4.75625
Tkred	1	1.47462
Tpersonal	1	2.98363
Tprogram	1	1.79625
TProj	1	3.74977
Smiljofl	1	2.92371
Sovrigt	1	1.16619
Sprodukt	1	2.60652
Ssocial	1	3.34628

Genomförandekanal lika med svenska organisationer och samarbetslandets organisationer samt typ av assistans lika med projektstöd har högst VIF. Men högsta $VIF_{\text{genomförande kanal svenska org.}} = 4{,}76$ ger, efter enkel manipulering av (3.22), att $R^2_{\text{genomförande kanal svenska org.}} = 0{,}79$ vilket inte är en så hög förklaringsgrad och de flesta variablerna ligger långt under den nivån. Därför kan vi utgå från att multicollinearity inte är ett påtagligt problem i det här datamaterialet. Det gäller oavsett modell som används.

5.2 Linear Dicriminant Analysis

Test av multivariat normalfördelning:

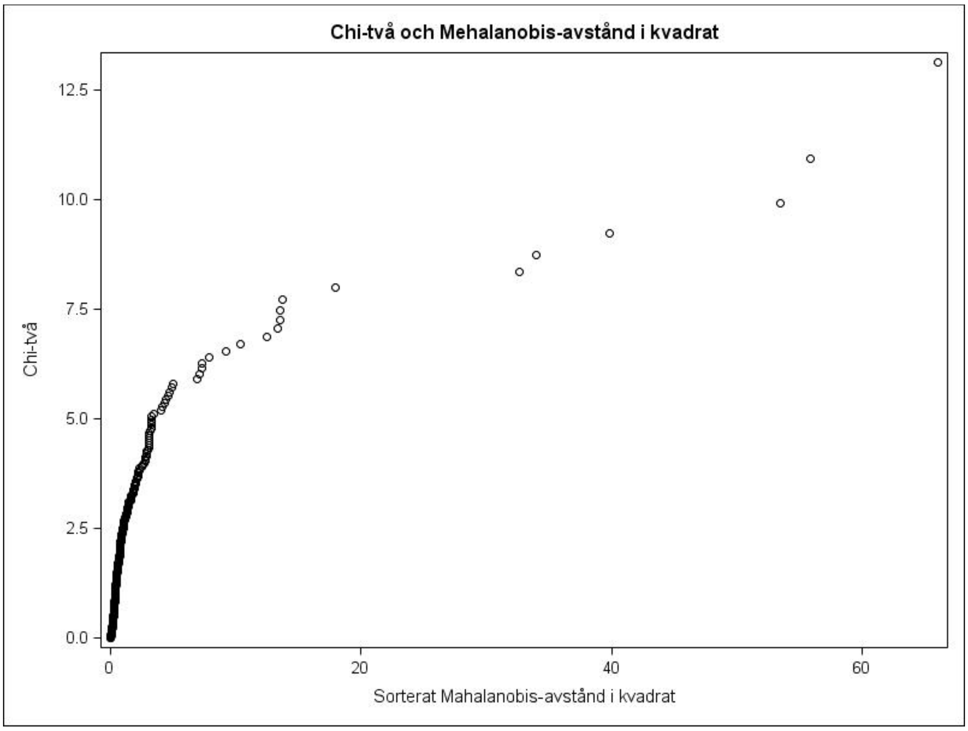

Bild2: Test av multivariat normalfördelning

Grafen visar att data inte är förenligt med antagandet om multivariat normalfördelning eftersom linjen inte är speciellt rak. Korrelationskoefficineten (som är ett när det linjära sambandet är perfekt) är 0,75. Emellertid ser det ut att finnas ett annat samband, nämligen att grafen växer som en logaritm-funktion. Därför provar jag att transformera variablerna insatslängd och insatsstorlek genom att logaritmera dem med basen 10. Det ger följande graf:

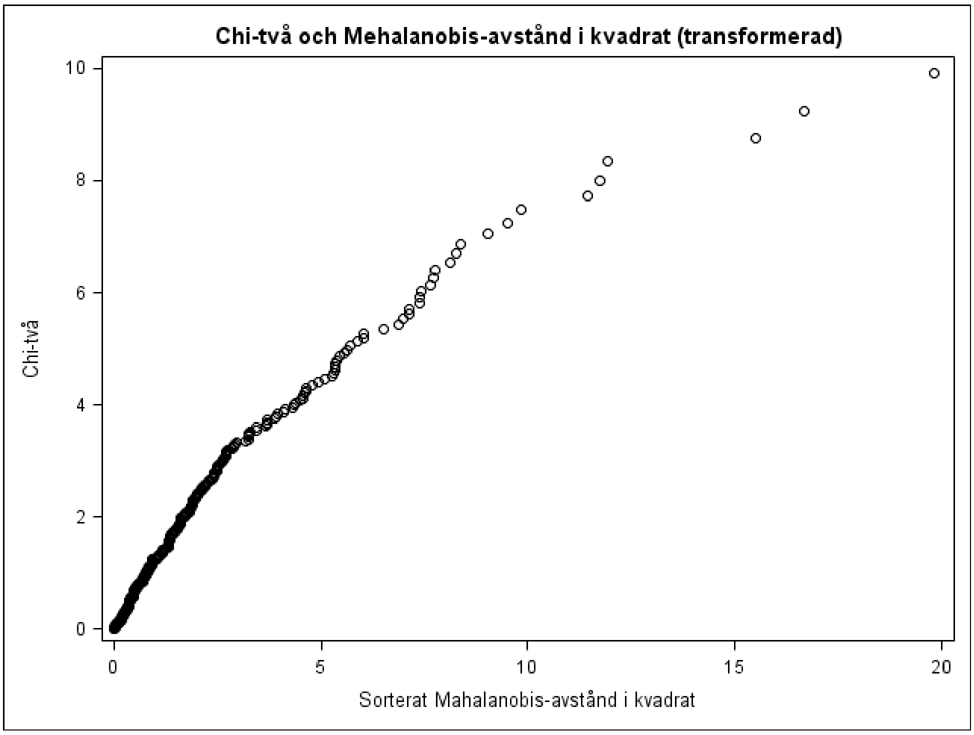

Bild3: Test av multivariat normalfördelning (transformerad)

Nu är sambandet i stort sätt linjärt och korrelationskoefficineten 0,96 vilket får anses vara tillfredsställande. Följande test undersöker ifall kovarians-matriserna är lika och nollhypotsen är att de är det, vilket är antagandet som helst ska vara uppfyllt

Test av homogenitet för inom kovariansmatriser

Chi-Square	DF	Pr > ChiSq
2.216593	3	0.5287

Tabell 2: Test av lika kovarians-matriser, LDA

Det går inte att förkasta hypotsen om lika kovarians och således är båda villkoren uppfyllda. Jag använder logaritmerade variablerna och undersöker ifall medelvärdena skiljer sig mellan grupperna:

Betyg = 0

Variable	N	Sum	Mean	Variance	Standard Deviation
loginsatslangd	180	310.01140	1.72229	0.02344	0.1531
loginsatsstorlek	180	170.92940	0.94961	0.19007	0.4360

Betyg = 1

Variable	N	Sum	Mean	Variance	Standard Deviation
loginsatslangd	175	292.54384	1.67168	0.02442	0.1563
loginsatsstorlek	175	166.46198	0.95121	0.15344	0.3917

Tabell 3: Test av lika variabel-medelvärden meller grupper, LDA

Medelvärdet för både logavtslängd och loginsatsstorlek är större för gruppen med betyg noll jämfört med betyg ett. Dock är skillnaden för varibeln loginsatsstorlek marginell. Ett formellt test av nollhypotsen att varib, lerna har lika medelvärden i båda grupperna ger:

Univariate Test Statistics

F-statistik, Tälj. DF=1, Nämn. DF=353

Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
loginsatslangd	0.1565	0.1547	0.0358	0.0262	0.0269	9.50	0.0022
loginsatsstorlek	0.4142	0.4147	0.001134	0.0000	0.0000	0.00	0.9710

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=175

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks Lambda	0.97025954	5.39	2	352	0.0049
Pillais spår	0.02974046	5.39	2	352	0.0049
Hotelling-Lawley-spår	0.03065206	5.39	2	352	0.0049
Roys största rot	0.03065206	5.39	2	352	0.0049

Tabell 4: Test av lika variabel-medelvärden aggregerat meller grupper, LDA

Det betyder att nollhypotsen för loginsatslängd, på <1 % signifikansnivå, kan förkastas medan vi inte kan förkasta att loginsatstorleken är lika mellan grupperna. Tillsammans är varibalenas medelvärden signifikant olika. Statistisk signifikans är för övrigt ett fenomen som påverkas mycket av antalet observationer och i det här fallet tycks den ”praktiska” signifikansen vara begränsad eftersom standdaravvikelserna gör att skillnaden i medelvärde mellan grupperna, på ungefär ett halvår $(10_0^{1,72}-10_1^{1,67})=(52,75-46,95)=5,80$, är liten i förhållande till den spridning som förekommer inom grupperna. Diskriminant-funktionen skattas:

Test of H0: The canonical correlations in

	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
	Eigenvalue	Difference	Proportion	Cumulative					
1	0.0307		1.0000	1.0000	0.97025954	5.39	2	352	0.0049

(Test of H0: The canonical correlations in the current row and all that follow are zero)

NOTE: The F statistic is exact.

DISCRIM-proceduren

Kanonisk diskriminant analys

Raw Canonical Coefficients

Variable	Can1
loginsatslangd	6.871965230
loginsatsstorlek	-0.895463200

Tabell 5: Diskriminant-funktion, LDA

Här framgår det att diskriminant-funktionen är signifikant och att loginsatslängd dominerar funktionen vilket är mycket rimligt med tanke på att bara den variabeln var påtagligt olika mellan grupperna. För att göra gruppindelningen levererar SAS inte poänggräns enligt ovan utan tillhandhåller istället två klassifikationsfunktioner vilkas koefficienter i sig inte har någon substantiell betydelse (Sharma 1996:256):

Variable	0	1
Konstant	-63.92327	-60.17638
loginsatslangd	75.67303	73.27333
loginsatsstorlek	-4.04625	-3.73356

Tabell 6: Funktioner för att bestämma gruppindelning, LDA

Om funktionen för grupp noll har högre ”poäng” så indelas observationen i den gruppen och om funktionen för grupp ett har poäng så indelas observationen i den gruppen. Det kanske viktigaste resultatet för LDA är hur många korrekta klassifikationer modellen gör:

Antal observationer och procent klassificerad till Betyg

From Betyg	0	1	Total
0	105	75	180
	58.33	41.67	100.00

1	76	99	175
	43.43	56.57	100.00
Total	181	174	355
	50.99	49.01	100.00

Tabell 7: Klassificeringsresultat, LDA

Totalt gör modellen $\frac{(105+99)}{355}=57{,}46\,\%$ korrekta prediktioner (totala antalet observationer är tre färre p.g.a. att det inte går att logaritmera 0); bättre än slumpen men inte så imponerande. Som tur är har vi ytterligare modeller att tillgå med många fler variabler.

5.2 Linear Probability Model

Först skattas modellen trots strukturella problemet med heterskedasticitet så skattningen av beroende varibeln kan användas som vikt enligt 3.20. I den viktade modellen, efter stepwise-proceduren, ingår följande variabler:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Modell	12	61.32207	5.11017	11.28	<.0001
Fel	331	149.93499	0.45298		
Korrigerad total	343	211.25706			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.89496	0.08549	49.64286	109.59	<.0001
Insatslangd	-0.00281	0.00131	2.06585	4.56	0.0334
internM	-0.26662	0.05726	9.82234	21.68	<.0001
internH	-0.57078	0.17593	4.76818	10.53	0.0013
internS	-0.38554	0.08841	8.61510	19.02	<.0001
externM	-0.11074	0.05744	1.68390	3.72	0.0547
externH	-0.23659	0.14516	1.20333	2.66	0.1041
externS	-0.26255	0.08657	4.16610	9.20	0.0026
GSamar	-0.09204	0.05974	1.07526	2.37	0.1243
Tkred	0.71836	0.19877	5.91675	13.06	0.0003
Tprogram	0.20370	0.10144	1.82658	4.03	0.0454
TProj	0.12131	0.05376	2.30679	5.09	0.0247
Smiljofl	-0.13886	0.06685	1.95448	4.31	0.0386

Bounds on condition number: 1.6343, 199.84

Alla variabler kvar i modellen är signifikanta på 0.1500-nivån.

Ingen annan variabel uppfyllde 0.1500-signifikansnivån för angivelse i modellen.

Vikt: Vikt_inv_kvadratrot_w

Tabell 8: Skattning av (viktad) LPM

Totalt används tio förklaringsvariabler och modellen är i sin helhet statistiskt signifikant på alla konventionella signifikansnivåer. Här är typ av assistans lika med kredit väldigt betydelsefull då förekomsten av den ökar sannolikheten, att en insats ska bedömas vara enligt planerna eller bättre, med i genomsnitt 71,83 %, relativt referenskategorin forskning, då alla andra variabler är oförändrade. Hög intern risk relativt låg sådan har betydande negativ effekt på sannolikheten. Det är intressant och mycket rimligt att alla risk-variabler har negativ effekt på sannolikheten, att insatsen ska bedömas enligt planerna eller bättre, relativt låg risk. Magnituden av den effekt som olika nivåer av högre risk har borde vara användbart för aktörer som arbetar med bistånd. Notera att insatslängd ingår även i den här modellen. Whites robusta standarderror:

		---Heteroscedasticity Consistent--			
Variable	DF	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.08577	10.43	<.0001	0
Insatslangd	1	0.00135	-2.09	0.0377	1.27489
internM	1	0.05872	-4.54	<.0001	1.53015
internH	1	0.17736	-3.22	0.0014	1.55083
internS	1	0.08292	-4.65	<.0001	1.40557
externM	1	0.05594	-1.98	0.0486	1.52091
externH	1	0.15469	-1.53	0.1271	1.63434
externS	1	0.08903	-2.95	0.0034	1.46483
GSamar	1	0.05377	-1.71	0.0879	1.18003
Tkred	1	0.14270	5.03	<.0001	1.19048
Tprogram	1	0.09740	2.09	0.0373	1.24648
TProj	1	0.05535	2.19	0.0291	1.34629
Smiljofl	1	0.06695	-2.07	0.0389	1.30854

Tabell 9: Whites robusta standarderror, LPM

Standardfelet förändras överlag mycket litet; ökar för några samt även minskar för nett par variabler. Praktikern bör utgå från de här avvikelserna när PLM tillämpas. Så hur bra är modellen på att predicera betyg korrekt d.v.s. hur ofta överensstämmer prediktionen med faktiskt betyg?

Overensstammelse	Frequency	Percent	Cumulative Frequency	Cumulative Percent

0	105	29.58	105	29.58
1	250	70.42	355	100.00

Tabell 10: Klassificeringsresultat, LPM

Variablen överensstämmelse har värdet ett när betyget har bedömts korrekt av modellen och är annars noll. Här är andelen rätt 70,42 %.

5.3 Logit Model

Stepwise-proceduren ger följande resultat:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood-förhå	93.3069	13	<.0001
Resultat	84.2747	13	<.0001
Wald	67.9777	13	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Insatslangd	1	3.9203	0.0477
Intern	3	25.1846	<.0001
Extern	3	8.8894	0.0308
Regleringsbrev	2	4.3940	0.1111
Typ_av_assistans	4	18.2368	0.0011

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	1.0435	0.5455	3.6586	0.0558	2.839
Insatslangd		1	-0.0139	0.00701	3.9203	0.0477	0.986
Intern	H	1	-3.4687	1.1614	8.9208	0.0028	0.031
Intern	M	1	-1.0797	0.2781	15.0776	0.0001	0.340
Intern	S	1	-1.6461	0.4774	11.8867	0.0006	0.193
Extern	H	1	-1.0566	0.7455	2.0087	0.1564	0.348
Extern	M	1	-0.5134	0.2897	3.1413	0.0763	0.598
Extern	S	1	-1.2459	0.4434	7.8945	0.0050	0.288
Regleringsbrev	K	1	1.1679	0.7788	2.2490	0.1337	3.215
Regleringsbrev	o	1	0.7050	0.4532	2.4195	0.1198	2.024
Typ_av_assistans	Proj	1	1.0949	0.3362	10.6062	0.0011	2.989
Typ_av_assistans	kred	1	3.7479	1.2308	9.2726	0.0023	42.430
Typ_av_assistans	personal	1	0.7981	0.4033	3.9161	0.0478	2.221
Typ_av_assistans	program	1	1.4637	0.5729	6.5276	0.0106	4.322

Tabell 12: Skattning av Logit

Modellen är mycket signifikant. Som framgår av ”Type 3 Analysis of Effects” inkluderas alla variabler som är signifikanta i sin helhet men det betyder inte att alla ”nivåer” är individuellt signifikanta. Till skillnad från LPM används inte dummy-varibaler utan klasser vilket mer är en teknisk aspekt än någonting av praktisk betydelse. Samma referenskategori gäller som i LPM. Hur som helst bör respektive nivå studeras var för sig och högt p-värde indikerar att variabeln inte bidrar mycket till att förklara variationen i beroende variabeln. Kolumnen längst till höger anger den effekt som variabeln (vid en enhets förändring då övriga variabler är oförändrade) har på oddskvoten relativt referenskategorin. När t.ex. regleringsbrev K (en klimatkonvention) är knuten till insatsen, istället för inget regleringsbrev, så är oddset att den ska vara enligt planerna eller bättre i genomsnitt 3,22 gånger så högt. Vidare så innebär ökning till medel intern risk, istället för låg, att oddset sjunker till 0,34 d.v.s. det är i genomsnitt ungefär en tredjedel så troligt att insatsen kommer att bedömas enligt planerna eller bättre när det förhållandet föreligger givet att allt annat lika. För den som vill veta hur sannolikheten påverkas kan 3.13 användas. Anta att förklaringsvariablerna implicerar en sannolikhet på 60 % och avtalslängden sedan ökar med sex månader då kommer sannolikheten att förändras så här:

$$\frac{dP_i}{dX_j} = \beta_j * P_i * (1 - P_i) = (-0{,}014 * 6) * 0{,}6(1 - 0{,}6) = -0{,}02$$

Alltså ungefär 2 %(-enheter). Whites robusta standarerror är inte tillgängligt och därför bör slutsatser dras med viss försiktighet. Andelen korrekta prediktioner:

Association of Predicted Probabilities and Observed Responses

Procent samstämmigt	77.9	Somers D	0.561
Procent ej samstämmigt	21.7	Gamma	0.564
Procent lika	0.4	Tau-a	0.281
Par	31500	c	0.781

Tabell 13: Klassificeringsresultat, Logit

Logit förutspår utfallet korrekt nästan åtta gånger av tio.

5.3 Probit Model

Stepwise-proceduren ger följande resultat:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood-förhå	93.4987	13	<.0001
Resultat	84.2922	13	<.0001
Wald	77.8433	13	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Insatslangd	1	3.9886	0.0458
Intern	3	27.3941	<.0001
Extern	3	9.3986	0.0244
Regleringsbrev	2	4.3373	0.1143
Typ_av_assistans	4	19.6349	0.0006

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.6050	0.3152	3.6852	0.0549
Insatslangd		1	-0.00789	0.00395	3.9886	0.0458
Intern	H	1	-2.0203	0.6067	11.0890	0.0009
Intern	M	1	-0.6693	0.1676	15.9483	<.0001
Intern	S	1	-0.9918	0.2791	12.6308	0.0004
Extern	H	1	-0.6595	0.4301	2.3519	0.1251
Extern	M	1	-0.3033	0.1738	3.0454	0.0810
Extern	S	1	-0.7506	0.2594	8.3726	0.0038
Regleringsbrev	K	1	0.6262	0.4332	2.0894	0.1483
Regleringsbrev	o	1	0.4196	0.2644	2.5195	0.1124
Typ_av_assistans	Proj	1	0.6775	0.1997	11.5093	0.0007
Typ_av_assistans	kred	1	2.2781	0.7144	10.1698	0.0014
Typ_av_assistans	personal	1	0.4881	0.2401	4.1311	0.0421
Typ_av_assistans	program	1	0.8854	0.3428	6.6711	0.0098

Association of Predicted Probabilities and Observed Responses

Procent samstämmigt	77.8	Somers D	0.559
Procent ej samstämmigt	21.8	Gamma	0.562
Procent lika	0.4	Tau-a	0.280
Par	31500	c	0.780

Tabell 14: Skattning av Probit och klassificeringsresultat

Modellen är mycket signifikant i sin helhet. I Probit ingår samma variabler som i Logit. Även andel korrekta prediktioner är nästan identisk. Det finns två skillander. Dels är inte koefficienterna jämförbara eftersom de påverkar sannolikheten genom olika modellstrukturer på olika sätt och dels saknas oddskvot när Probit används vilket gör tolkningen av parametrarna problematisk. Den påverkan förändring i en förklaringsvaribel har på sannolikheten beräknas genom att sannolikhetsnivån före förändrigen jämförs med nivån efteråt.

5.4 Multiple-Group Linear Discriminant Analysis

Här är grupp $0=U, 1=A, 2=G\, och\, 3=VG$. Nu är dessvärre kovarian-matiserna inte lika mellan grupperna:

Test av homogenitet för inom kovariansmatriser

Chi-Square	DF	Pr > ChiSq
20.639892	9	0.0143

Tabell 15: Test av lika kovarians-matriser, MLDA

Nollhypotsen om lika kovarians förkastas vid konventionella signifikansnivåer. Det betyder att signifikanstest av diskriminant-funktionerna och klassifiering för göras med försiktighet.

Univariate Test Statistics

F-statistik, Tälj. DF=3, Nämn. DF=351

Variable	Total Standard Deviation	Pooled Standard Deviation	Between Standard Deviation	R-Square	R-Square / (1-RSq)	F Value	Pr > F
loginsatslangd	0.1565	0.1550	0.0303	0.0282	0.0290	3.39	0.0182
loginsatsstorlek	0.4142	0.4150	0.0312	0.0043	0.0043	0.50	0.6813

Tabell 16: Test av lika variabel-medelvärden meller grupper, MLDA

Medelvärdet för loginsatslängd är åter statistiskt signifikant olika mellan fyra grupperna. Nu förekommer två diskriminant-funktioner:

DISCRIM-proceduren

Kanonisk diskriminant analys

Raw Canonical Coefficients

Variable	Can1	Can2
loginsatslangd	6.856651356	0.222761041
loginsatsstorlek	-0.946480972	2.379933763

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.0329	0.0286	0.8854	0.8854	0.96404334	2.16	6	700	0.0454
2	0.0043		0.1146	1.0000	0.99576026	0.75	2	351	0.4744

Tabell 17: Diskriminant-funktioner, MLDA

Den första domineras av loginsatlängd medan den andra är mer viktad mot loginsatsstorlek. Det är rimligt eftersom poängen för första och andra diskriminant-funktionen inte ska korrelera. Dock är bara första diskriminant-funktionen signifikant. Modellens kapacitet att gissa rätt:

Antal observationer och procent klassificerad till Betyg

From Betyg	0	1	2	3	Total
0	3	16	11	0	30
	10.00	53.33	36.67	0.00	100.00
1	1	76	73	0	150
	0.67	50.67	48.67	0.00	100.00
2	0	59	91	3	153
	0.00	38.56	59.48	1.96	100.00
3	0	6	15	1	22
	0.00	27.27	68.18	4.55	100.00
Total	4	157	190	4	355
	1.13	44.23	53.52	1.13	100.00

Tabell 18: Klassificeringsresultat, MLDA

Modellen gör rätt gruppindelning $\frac{(3+76+91+1)}{355}=48{,}17\%$ av gångerna vilket är bättre än slumpen som blir 25 % ifall insatser delas in helt slumpvis i grupperna.

5.5 Polytomous Logistic Regression

Stepwise-proceduren ger följande resultat:

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood-förhå	115.7155	18	<.0001
Resultat	140.4406	18	<.0001
Wald	91.7129	18	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Intern	9	51.7974	<.0001
Extern	9	19.1265	0.0241

Analysis of Maximum Likelihood Estimates

Parameter		Betyg	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		3	1	1.6748	0.6995	5.7331	0.0166	5.338
Intercept		2	1	3.1210	0.6406	23.7336	<.0001	22.669
Intercept		1	1	1.6798	0.6683	6.3181	0.0120	5.364
Intern	H	3	1	-22.9971	14711.3	0.0000	0.9988	0.000
Intern	H	2	1	-4.8712	1.3293	13.4280	0.0002	0.008
Intern	H	1	1	-2.9245	1.0299	8.0626	0.0045	0.054
Intern	M	3	1	-1.1837	0.8220	2.0738	0.1498	0.306
Intern	M	2	1	-1.2933	0.7136	3.2848	0.0699	0.274
Intern	M	1	1	0.0643	0.7250	0.0079	0.9293	1.066
Intern	S	3	1	-16.6075	689.6	0.0006	0.9808	0.000
Intern	S	2	1	-3.1609	0.7924	15.9130	<.0001	0.042
Intern	S	1	1	-1.8698	0.7786	5.7672	0.0163	0.154
Extern	H	3	1	0.2655	1.4378	0.0341	0.8535	1.304
Extern	H	2	1	-0.4852	1.0710	0.2052	0.6505	0.616
Extern	H	1	1	0.6629	0.9373	0.5002	0.4794	1.940
Extern	M	3	1	-0.5100	0.7816	0.4257	0.5141	0.601
Extern	M	2	1	0.6735	0.6188	1.1846	0.2764	1.961
Extern	M	1	1	1.3077	0.6244	4.3860	0.0362	3.698
Extern	S	3	1	-0.9797	1.0005	0.9589	0.3275	0.375
Extern	S	2	1	-0.8303	0.6948	1.4282	0.2321	0.436
Extern	S	1	1	0.1629	0.6680	0.0595	0.8073	1.177

Tabell 18: Skattning av PLR

Modellen är mycket signifikant. Som resultatet visar är interceptet och övriga koefficienter olika för respektive logit-funktion. Betygen är kodade på samma sätt som för MLDA och referensbetyget är noll d.v.s. U. Endast intern- och extern risk är signifikanta varav enskilda nivåer på variablerna ofta har höga p-värden. För de flest skattningarna gäller att högre risk leder till lägre betyg t.ex. så har intern hög risk lägre odds relativt låg risk oavsett vilket betyg som jämförs med U.

5.6 Proportional Odds Model

Till att börja med är det nödvändigt att undersöka huruvida modellens antagande att regressionskoefficinterna är samma oavsett ordinal kategori (se 3.17). Ett score-test ger följande resultat

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
73.8705	22	<.0001

Tabell 18: Score-test, POM

Nollhypotsen är att koefficinterna (förutom interceptet) är lika. Eftersom nollhypotsen förkastas vid alla konventionella signifikansnivåer betyder det att antagndet inte är uppfyllt och modellens giltighet bör ifrågasättas. Förvisso finns det modeller som behandlar beroende variabelns ordinalskla utan att ställa ovan villkor bl.a. ”Continuation Ratio Model” (O'Connel 2006:54-57) men eftersom det förekommer förhållandevis få U och VG så är det inte värt att fördjupa analysen ytterligare. Här redovisas dock skattningen av POM men slutsatser riskerar alltså att vara felaktiga:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Insatslangd	1	3.6165	0.0572
Intern	3	43.7081	<.0001
Extern	3	9.0106	0.0292
Typ_av_assistans	4	16.7569	0.0022

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	3	1	-1.9127	0.4895	15.2657	<.0001	0.148
Intercept	2	1	1.2243	0.4741	6.6689	0.0098	3.402
Intercept	1	1	4.2906	0.5344	64.4712	<.0001	73.010
Insatslangd		1	-0.0113	0.00594	3.6165	0.0572	0.989

Intern	H	1	-4.1541	0.7872	27.8485	<.0001	0.016
Intern	M	1	-0.9499	0.2567	13.6926	0.0002	0.387
Intern	S	1	-2.2007	0.4435	24.6217	<.0001	0.111
Extern	H	1	-0.6800	0.6519	1.0881	0.2969	0.507
Extern	M	1	-0.5023	0.2615	3.6893	0.0548	0.605
Extern	S	1	-1.1645	0.4001	8.4734	0.0036	0.312
Typ_av_assistans	Proj	1	0.8307	0.2885	8.2916	0.0040	2.295
Typ_av_assistans	kred	1	2.7685	0.9250	8.9581	0.0028	15.934
Typ_av_assistans	personal	1	0.2964	0.3478	0.7262	0.3941	1.345
Typ_av_assistans	program	1	0.9763	0.4846	4.0587	0.0439	2.655

Tabell 19: Skattning av POM

Notera att endast interceptet är olika mellan (ordinala) katagorierna och att stepwise-proceduren leder till att samma variabler ingår i POM som i Logit. Andelen korrekta klassifikationer är något lägre jämfört med Logit och Probit:

Association of Predicted Probabilities and Observed Responses

Procent samstämmigt	75.2	Somers D	0.510
Procent ej samstämmigt	24.2	Gamma	0.513
Procent lika	0.5	Tau-a	0.320
Par	39366	c	0.755

Tabell 20: Klassificeringsresultat POM

6: Slutsats

Logit lämpar sig bäst för att beräkna förväntade sannolikheten att en insats ska uppfylla uppsatta mål eller bättre. Jämfört med Probit, som är lika bra på att klassifiera insatser, så är det betydligt smidigare att använda Logit för att utröna förklaringsvariblernas påverkan i och med att förändrade odds kan användas. LPM har också en del förtjänster, i synnerhet regressionskoeffienternas direkta samband med sannolikhet och robusta standardavikelser som gör att det går att ta hänsyn till eventuell heteroskedasticitet. LDA uppmärksammar insatslängdens betydelse men överlag är multivariata metoderna av begränsad betydelse i det här sammanhanget så länge förklaringsvariablerna framförallt är av kvalitativ natur. PLR har viss behållning och kan ge vägledning angående den påverkan förklaringsvariabler har på sannolikheten när alla fyra betyg används som beroende variabel. POM har visat sig vara icke tillämpbar.

Av varibalerna i Logit bidrar insatslängd, intern risk, extern risk och typ av assistans signifikant till förklarningen av utfallet på beroende variabeln (se *tabell 12*). Varken policymarkörerna, genomförandeplats eller sektor bidrog tillräckligt mycket för att bli inkluderade i modellen. Av alla 358 insatser som analyserats predicerar modellen rätt utfall 279 gånger (77,9 %). Så biståndspratikern kan med fördel använda modellen innan resurser fördelas. För att illustrera tillämpaningen har jag valt ut tre stycken insatser slumpmässigt i datan (här har modellen rätt två gånger och är felaktig en gång):

Insats-ID	Insatslängd	Intern risk	Exten risk	Reglerings brev	Typ av assistans	Bedömd sannolikhet och betyg	Faktiskt betyg
76003104	49	S	S	Nej	Projektstöd	19,24 % < 50 % $\rightarrow A/U$	A/U
76002820	76	M	L	Nej	Projektstöd	37,49 % < 50 % $\rightarrow A/U$	A/U
72600419	44	M	M	Nej	Programstöd	69,34 % > 50 % $\rightarrow G/VG$	A/U

Sannolikheten för t.ex. 76003104 beräknas på följande sätt med hjälp av 3.12 och skattade koefficienterna i *tabell 12*:

$$P(Y_{76003104}=1)=\frac{1}{[1+\exp(-\tilde{X}_{76003104}{}'*\tilde{\beta})]}$$

$$=\frac{1}{1+\exp[-(1{,}0435+49*-0{,}0139+1*-1{,}6461+1*-1{,}2459+0+1*1{,}0949)]}=19{,}24\ \%$$

1,0435 är skattningen av interceptet, -0,0139 är skattningen av regressionskoefficienten för insatslängd varav 49 är antal månader, -1,6461 är intern risk lika med S (Substantial) relativt L (Low), -1,2459 är extern risk S relativt L, 0 uttrycker att regleringsbrev saknas som är referenskategorin och 1,0949 är typ av assistans lika med projektstöd relativt forskning.

Om vi återknyter till diskussionen i inledningen så anger data att insats 76003104 kostar 75 000 000 SEK. Låt oss anta att den beräknade sannolikheten motsvarar hur stor andel av uppsatta målen som insatsen kan förväntas uppfylla (varav 50 % anger gränsvärdet då insatsens betyg börjar bedömas som enligt planerna eller bättre). Om vi vidare antar att insatsens värde är proportionerlig mot graden av måluppfyllelse så går det att beräkna exakt hur mycket insatsen, vid ett bästa scenario, minst måste vara värd för att förväntade nyttan U ska vara större än kostanden:

75 000 000 = U * 19,24%, U = 389 812 889 SEK.

Försiktighet bör samtidigt iakttas vid en sådan tolkning då modellen inte gör anspråk på att uttrycka graden av måluppfyllelse utan enbart vilket betyg vi kan förvänta oss givet förklaringsvariablerna. En mer restriktiv applicering är att biståndshandläggaren bara genomför insatser som förväntas vara enligt planerna eller bättre (förutom då nyttan överstiger kostnaden även när insatsen inte är framgångsrik) samt sätter mer tilltro till insatser med hög sannolikhet. Sedan kan han eller hon t.ex. fråga sig om det är värt att bedriva en insats med hög extern risk istället för låg sådan trots att det är förknippat med ett odds som gör det en tredjedel så troligt att insatsen ska gå enligt planerna eller bättre.

Referenslista

- Gujarati, D. & Porter, D. 2009, *Basic Econometrics*. New York: McGraw-Hill.
- Johnson, R. & Wichern, D. 2007, *Applied Multivariate Statistical Analysis*. New Jersey: Pearson Education.
- Johnson, D., Kupper, L., Nizam, A. & Muller, K. 2008, *Applied Regression Analysis and other Multivariate Methods*. California: Brooks/Cole.
- JO'Connel, A. 2006, *Logistic Regression Models for Ordinal Response Variables*. California: Sage Publications.
- SAS (online). Tillgänglig från http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect042.htm (besökt 22 december 2011)
- Sharma, S. 1996, *Applied Multivariate Techniques*. Canada: John Wiley & Sons.
- Sida (online). Tillgänglig från http://www.sida.se/Svenska/Om-oss/Budget/ (besökt 22 december 2011)
- Stock, J. & Watson, M. 2007, *Introduction to Econometrics*. Boston: Pearson Education.
- Wackerly, D., William, M. & Scheaffer, R. 2008, *Mathematical Statistics with applications*. Boston: Pearson Education.

Appendix

A.1 Datamaterial redovisas i separat bilaga.

A.2 Minstakvadratmetoden.

Metoden bygger på att summan av alla kvadrerade avstånd mellan skattade observation $\hat{Y}_i$ och faktiska observation Y_i minimeras.

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{b}_o-\hat{b}_1*X_i)^2 \quad \text{(A.1)}$$

β_o och β_1 är de värden på $\hat{b}_o$ och $\hat{b}_1$ som åstadkommer minimeringen. Genom att derivera A.1 med avseende på $\hat{b}_o$ och $\hat{b}_1$ och sätta uttrycket lika med noll kan β_o och β_o identifieras.

$$\frac{d}{d\,b_o}\sum_{i=1}^{n}(Y_i-\hat{b}_o-\hat{b}_1*X_i)^2=-2*\sum_{i=1}^{n}(Y_i-\hat{b}_o-\hat{b}_1*X_i)=0$$

$$\frac{d}{d\,b_1}\sum_{i=1}^{n}(Y_i-\hat{b}_o-\hat{b}_1*X_i)^2=-2*\sum_{i=1}^{n}(Y_i-\hat{b}_o-\hat{b}_1*X_i)*X_i=0$$

Båda uttrycken divideras med *n* och det ger att:

$$\bar{Y}-\hat{\beta}_o-\hat{\beta}_1*\bar{X}=0$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i*Y_i-\hat{\beta}_0*\bar{X}-\frac{\hat{\beta}_1}{n}\sum_{i=1}^{n}X_i^2=0$$

De båda kombineras och skattningarna av koefficienterna beräknas:

$$\beta_1=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \quad \text{(A.2)}$$

$\beta_0 = \bar{Y} - \beta_1 * \bar{X}$ (A.3)

Metoden kan generaliseras till fler variabler och då sätts partiella derivata lika med noll för varje variabel och koefficienterna kan lösas ut enligt samma princip.

A.3 Maximum Likelihood Estimate Logit

Binära Y_i är följer en en Bernoulli-sannolikhetsfördelning. Sannolikheten att $Y_i = 1$ är P_i och $Y_i = 0$ har sannolikheten $(1 - P_i)$. Bernoulli-variabeln har följande sannolikhetsfördelning:

$$P(Y_i | P) = P^{Y\_i} * (1 - P)^{1-Y\_i}, Y_i = 0,1 \quad \text{(A.4)}$$

Alla $Y_{1,} Y_{2,} \ldots, Y_n$ är oberoende av varandra. Sannolikheten att observera $Y_{1,} Y_{2,} \ldots, Y_n$ givet P_i är produkten av alla sannolikheter:

$$L(Y_{1,} Y_{2,} \ldots, Y_n | P) = P^{Y\_1} * (1 - P)^{1-Y\_1} * P^{Y\_2} * (1 - P)^{1-Y\_2} * \ldots * P^{Y\_n} * (1 - P)^{1-Y\_n}$$
$$= \prod_{i=1}^{n} P^{Y\_i} * (1 - P)^{1-Y\_i} \quad \text{(A.5)}$$

där *L* är likeliehood-funktionen. Anta att $Y_i = 1$ för första n_1 observationerna och att $Y_i = 0$ för övriga observationer så att:

$$L = \prod_{i=1}^{n_1} P * \prod_{i=n\_1+1}^{n} (1 - P) \quad \text{(A.6)}$$

P_i beror i sin tur på förklaringsvariablerna $\tilde{X}_i$ på det sätt sannolikhetsfunktionen för Logit specificerar. Då blir *L*:

$$L = \frac{\prod_{i=1}^{n_1} \exp(\tilde{X}_i{}' * \tilde{\beta})}{\prod_{i=1}^{n} [1 + \exp(\tilde{X}_i{}' * \tilde{\beta})]} \quad \text{(A.7)}$$

Därefter gäller det att identifiera de koefficienter som maximerar *L* så att sannolikheten att observera $Y_{1,} Y_{2,} \ldots , Y_n$ blir så stor som möjligt. Statistisk programvara t.ex. SAS utför den operationen.

A.3 Maximum Likelihood Estimate Probit

För att identifiera koefficientern i Probit sätts $P_i = \Phi(\tilde{X}_i{}' * \tilde{\beta})$ in i A.6 istället för sannolikhetsfunktionen för Logit.