Grundläggande om *logistisk regression*

Här diskuteras några grundläggande idéer om så kallad *logistisk regression*. Det finns mängder av information tillgänglig både på nätet och i otaliga böcker om olika typer av regressionsanalys.

Rubriker:

1. Vad är 'vanlig regression'?

Vanlig regression innebär att man försöker hitta ett matematiskt samband mellan en kontinuerlig *Y*-variabel och en eller flera förklaringsvariabler som oftast kallas *X*-variabler. Man vill alltså undersöka om och hur dessa påverkar *Y*-resultatet.

Antal att man vill titta på t.ex. tillväxt av ett ytskikt och se hur detta påverkas av olika faktorer. Grundat på tidigare kunskap och några idéer väljer man ut ett antal dylika faktorer som skall användas. Man behöver också bestämma nivåer på varje faktor (minst två nivåer, väljer man fler blir mätningarna mer komplicerade och dyrare). (Allt detta behandlas inom ämnet *Försöksplanering*.)

Om man studerar någon viktig mätning i människokroppen kan man oftast inte genomföra detta som ett experiment utan man hänvisas till data samlad på annat sätt.

En typisk regressionsmodell kan se så här:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

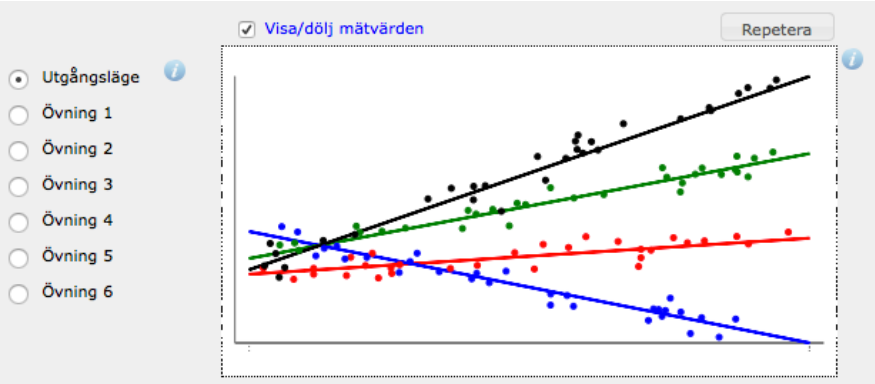
Här representerar sista termen ('epsilon') alla de variabler – kända eller okända – som inte ingår i modellen men som ändå påverkar resultatet. Analysen går ut på finna värden på modellens parametrar, de tre 'beta'-koefficienterna. Om en koefficient är noll eller nära noll innebär det att X-variabeln inte tycks påverka mätresultatet och kan då uteslutas ur modellen.

Analysen kan också på andra sätt avslöja om man har valt fel modell, kanske det finns krökta samband av mer eller mindre komplicerad form.

('Vanlig regression' kallas ibland på engelska för *Ordinary Least Square* (*OLS*) eftersom den bakomliggande matematiken kallas på svenska för *minsta-kvadrat-metoden*.)

Ett exempel från ovn.ing-stat.se

Här finns det en (1) kontinuerlig X-variabel samt en X-variabel som består av fyra olika grupper (symboliserade av fyra olika färger).



2. Vad är 'logistisk regression'?

I väldigt många situationer studerar man proportioner typ *felkvot*, *dödlighet*, *överlevnad*, etc. Då har man en klar definition för varje observerat fall dvs varje enhet som studeras. En detalj är antingen 'godkänd' eller 'ej godkänd', eller en person 'avled av sjukdom Z' eller inte.

Man har alltså ett antal sådan observationer och från detta beräknar man en proportion (ibland används ordet *kvot* som i 'felkvot'.). Precis som vid 'vanlig regression' vill man förklara proportionen med ett antal olika förklaringsvariabler.

Vid studier på människor är vanliga förklaringsvariabler *kön*, *ålder*, *vikt*, *matvanor*, etc, etc. (Förvånande nog är denna situation svårare rent matematiskt och kräver ett datorprogram som kan göra en iterativ beräkning. I 'vanlig regression' finns det i stället ett antal formler som man kan avända med en enkel kalkylator.)

Logistisk modell. En vanlig modell för att analysera proportioner är en s.k. logistisk modell där proportionen (p) är en funktion av ett antal förklaringsvariabler (X) enligt nedanstående uttryck:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...)}$$

Det simulerade exemplet med kvalitetsförbättring använder bara en förklaringsvariabel (*tid, T*) i modellen:

$$p = \frac{\exp(\beta_0 + \beta_1 T)}{1 + \exp(\beta_0 + \beta_1 T)}$$

3. Beräkning av modellens parametrar

I modellen för kvalitetsförbättring finns det två obekanta, nämligen 'beta0' och 'beta1', dvs modellens två parametrar (koefficienter). Om vi då har två punkter på kurvan kan vi lösa ut dessa två parametrar och få två ekvationer. Vi kallar dessa två punkter (p_0, t_0) samt (p_1, t_1)

Med hjälp av lite matematisk manipulation erhåller följande två uttryck:

$$\beta_0 = \ln(p_0) - \ln(1 - p_0) - \beta_1 t_0 \qquad \beta_1 = \frac{\ln(p_1) - \ln(1 - p_1) + \ln(1 - p_0)}{t_1 - t_0}$$

Antag att kurvan startar vid felkvot 0.10 (10%) vid tidpunkt t = 0 och slutar vid felkvot 0.01 (1%) vid tidpunkt t = 365, dvs ett år senare. Genom att först beräkna 'beta1' och sedan 'beta0' erhålles numeriska värden och därefter kan kurvan ritas upp (blå linje på skärmen).

Kurvan visar då förväntad felkvot vid varje tillfälle om kvalitetsförbättringarna följer det uppsatta målet.

4. Bestämning av parametrar ur data

För att kontrollera att förbättringsarbetet följer det uppsatta målet, måste samma parametrar beräknas ur data. Dessa innehåller dock en slumpkomponent och sålunda blir beräkningen av dessa inte exakt lika modellens parametrar.

5. Övrigt

I programvaror för regressionsanalys finns det en mängd metoder och hjälpmedel för att bl.a. studera hur den ur data skattade modellen överensstämmer med den planerade modellen.