

Statistisk analys av komplexa data

Kategoriska data, ht 2017

Bertil Wegmann

STIMA, IDA, Linköpings universitet

■ Kategoriska data

■ oordnade data (nominal)

- “ja” eller “nej”, vilken yrkesgrupp en individ tillhör, en kunds val av tvättmedel

■ ordnade data (ordinal)

- betyg, preferenser för olika frågor

■ **Korstabeller:** goodness-of-fit test för att testa om samband föreligger mellan variabler som antar olika kategorier

■ **Multinomialfördelning:** generalisering av binomialfördelning

■ ***Qualitative response (QR) models:*** beroende variabeln i en regressionsmodell antar diskreta utfall

- Logit och probit modeller: binär beroende variabel
- Multinomial logit modell: beroende variabeln antar flera oordnade kategorier
- Ordad logit och probit modell: beroende variabeln antar flera ordnade kategorier

- En nationell rapporteringsservice för bränder har konstaterat bland bostadsbränder att 73 % inträffar i familjehem, 20 % i lägenheter och 7 % i andra typ av bostäder.
 - Frågeställning: vad är sannolikheten en slumpmässigt vald dag att 2 bränder är i familjehus, 1 är i en lägenhet och 1 i en annan typ av bostad? (Multinomialfördelning)
- Försäljning av ett samlarmynt på en auktion: försäljning sker om det högsta budet blir högre än säljarens hemliga reservationspris. Kodas med värdet 1 vid försäljning och 0 vid ingen försäljning.
 - Frågeställning: hur påverkas sannolikheten för försäljning av specifika auktionsegenskaper? (Logit/Probit modell)
- Konsumenters val av tandkräm i en livsmedelsbutik en viss dag.
 - Frågeställning: vad beror val av tandkräm på? Ålder? Kön? Utbildning? Lön? TV-tittande? (Multinomial logit modell)

- Inställning om en utbyggnad av kommunens tennishall. Låt 5 vara “mycket positiv”, 4 “positiv”, 3 “neutral”, 2 “negativ”, 1 “mycket negativ”. Inställningarna kan rankas gentemot varandra, men skillnaderna mellan nivåerna behöver inte vara samma.
 - Frågeställning: kan grad av inställning förklaras med hjälp av bakgrundsvariabler för kommunens invånare? (Ordnad logit/probit modell)
- Betyg på en kurs. Kategorierna VG, G och U kan kodas till numeriska värden och rankas gentemot varandra.
 - Frågeställning: vad beror betyget på en kurs av? Tidigare kursbetyg? Förkunskaper? Intelligens? Ålder? Akademisk ålder? (Ordnad logit/probit modell)

- **Exempel:** test av nollhypotes om inget samband mellan kön och flygplansreservationer.

	Kvinna	Man	TOTALT
Resebyrå	256	74	330
Internet	41	42	83
Telefon	66	34	100
TOTALT	363	150	513

- Urvalet om n st. individer har kors-klassificerats inom $r = 3$ st. olika kategorier av flygplansreservationer och $k = 2$ kategorier för kön. Under nollhypotesen förväntar vi oss inget samband mellan dessa egenskaper. Antalet förväntade observationer i varje cell (i, j) under nollhypotesen ges av

$$E_{ij} = \frac{R_i K_j}{n},$$

där R_i och K_j är antalet observationer i respektive i :te rad och j :te kolumn.

- Ett goodness-of-fit test ger att nollhypotesen förkastas om

$$\sum_{i=1}^r \sum_{j=1}^k \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)(k-1), \alpha}$$

där O_{ij} är den observerade frekvensen i cell (i, j) .

- Enligt data från tabellen förkastas nollhypotesen på alla rimliga signifikansnivåer, eftersom $\chi^2_{obs} = 26.8 > \chi^2_{(3-1)(2-1), 0.005} = 10.60$. Alltså finns det stöd i data att kön och flygplansreservationer är associerade med varandra.

- Generaliserar binomialfördelningen för 2 st. kategorier till $k > 2$ st. kategorier.
- Om man känner till sannolikheterna $(p_1, p_2, \dots, p_k)$ för de olika k kategorierna, så ges sannolikheten från multinomialfördelningen som

$$P(x_1, x_2, \dots, x_k) = \binom{n}{x_1, x_2, \dots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

- En nationell rapporteringsservice för bränder har konstaterat bland bostadsbränder att 73 % inträffar i familjehem, 20 % i lägenheter och 7 % i andra typ av bostäder.
 - Frågeställning: vad är sannolikheten en slumpmässigt vald dag att 2 bränder är i familjehus, 1 är i en lägenhet och 1 i en annan typ av bostad?

- Den beroende variabeln Y är binär, d.v.s. antingen är $Y = 0$ eller $Y = 1$. Vi vill skatta hur det förväntade värdet på Y beror på olika förklaringsvariabler i en regressionsmodell.
- Det förväntade värdet för Y är $E[Y] = P(Y = 1)$. **Exempel:** hur påverkas sannolikheten för försäljning, $P(Y = 1)$, av specifika auktionsegenskaper?
- Vi vill alltså förklara sannolikheten för den ena kategorin utifrån förklaringsvariablerna i regressionsmodellen. Detta kan uppnås med hjälp av olika funktioner, $F(\beta\mathbf{x}')$, av parametervektorn $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)$ till vektorn med förklaringsvariabler $\mathbf{x} = (1, x_1, x_2, \dots, x_k)$, d.v.s.

$$P(Y = 1) = F(\beta\mathbf{x}')$$

- Linjär regressionsmodell, $F(\beta\mathbf{x}') = \beta\mathbf{x}'$, är inte lämplig för detta ändamål, eftersom det måste gälla att $0 \leq \beta\mathbf{x}' \leq 1$.

- Normalfördelningen ger upphov till **probit modellen** där

$$P(Y = 1) = \int_{-\infty}^{\beta \mathbf{x}'} \phi(t) dt = \Phi(\beta \mathbf{x}'),$$

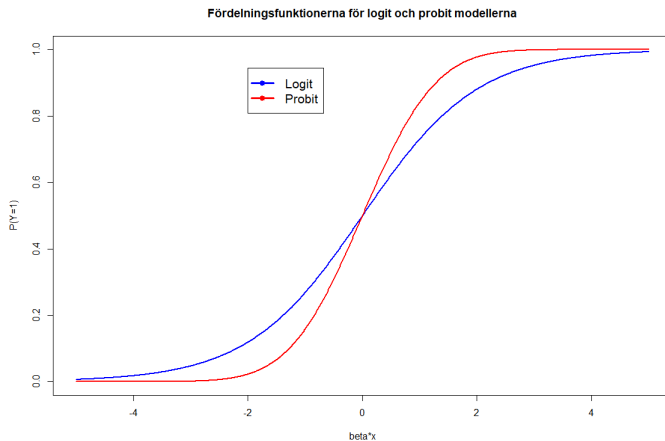
där $\Phi(\cdot)$ är fördelningsfunktionen för standard normalfördelningen.

- **Logit modellen** kommer från log-oddset som en linjärkombination av förklaringsvariablerna, $\ln \left(\frac{P(Y=1)}{P(Y=0)} \right) = \ln \left(\frac{P(Y=1)}{1-P(Y=1)} \right) = \beta \mathbf{x}'$, vilket ger

$$P(Y = 1) = \frac{e^{\beta \mathbf{x}'}}{1 + e^{\beta \mathbf{x}'}} = \Lambda(\beta \mathbf{x}'),$$

där $\Lambda(\cdot)$ är den logistiska fördelningsfunktionen.

- Den logistiska fördelningsfunktionen liknar fördelningsfunktionen för standard normalfördelningen, förutom i svansarna nära sannolikheterna 0 och 1 där den är tjockare.



- $Y_i \stackrel{iid}{\sim} \text{Bernoulli} (\theta = P(Y_i = 1) = F(\beta \mathbf{x}_i'))$. Likelihoodfunktionen ges som

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n | \theta) = \prod_{y_i=0} P(Y_i = 0) \prod_{y_i=1} P(Y_i = 1)$$

$$= \prod_{y_i=0} [1 - F(\beta \mathbf{x}_i')] \prod_{y_i=1} F(\beta \mathbf{x}_i') = \prod_{i=1}^n [1 - F(\beta \mathbf{x}_i')]^{1-y_i} [F(\beta \mathbf{x}_i')]^{y_i}$$

- Log Likelihoodfunktionen blir då

$$\ln L = \sum_{i=1}^n [y_i \ln F(\beta \mathbf{x}_i') + (1 - y_i) \ln (1 - F(\beta \mathbf{x}_i'))]$$

- Maximum likelihood skattningen fås genom att maximera log-likelihooden numeriskt.

- I logit modellen är β_j den marginella effekten på log-oddset från respektive förklaringsvariabel x_j , givet att $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k)$ hålls konstanta. Marginell effekt på log-oddset är svårtolkat!
- Oddset förväntas förändras med en faktor $e^{m\beta_j}$ då en förklaringsvariabel x_j ökar med m enheter, givet att $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k)$ hålls konstanta.

- Vad är den marginella effekten på $P(Y = 1) = E[y]$? Generellt,

$$\frac{\partial E[y]}{\partial x_j} = \left\{ \frac{dF(\beta \mathbf{x}')}{d(\beta \mathbf{x}')} \right\} \beta_j = f(\beta \mathbf{x}') \beta_j$$

- Logit modellen:

$$\frac{\partial E[y]}{\partial x_j} = \left\{ \frac{d\Lambda(\beta \mathbf{x}')}{d(\beta \mathbf{x}')} \right\} \beta_j = \frac{e^{\beta \mathbf{x}'}}{(1 + e^{\beta \mathbf{x}'})^2} \beta_j = \Lambda(\beta \mathbf{x}') (1 - \Lambda(\beta \mathbf{x}')) \beta_j$$

- Probit modellen:

$$\frac{\partial E[y]}{\partial x_j} = \left\{ \frac{d\Phi(\beta \mathbf{x}')}{d(\beta \mathbf{x}')} \right\} \beta_j = \phi(\beta \mathbf{x}') \beta_j$$

Exempel: skattning av logit och probit modellerna

- Spector och Mazzeo (1980) analyserade effekten av en ny undervisningsmetod i nationalekonomi. Datamaterialet bestod av
 - beroende variabel GRADE, en indikator med värdet 1 om studentens betyg förbättrades och 0 annars.
 - förklaringsvariabel GPA: genomsnittsbetyg för varje individ i urvalet.
 - förklaringsvariabel TUCE: poäng på ett test om förkunskaper.
 - förklaringsvariabel PSI antar värdet 1 om den nya undervisningsmetoden användes och 0 annars.
- Skattning av probit och logit modellerna kan utföras m.h.a. statistiska programvaran *R*:

```
myl<-glm(GRADE~GPA+TUCE+PSI,family="binomial")
```

```
summary(myl)
```

```
myp<-glm(GRADE~GPA+TUCE+PSI,family="binomial"(link="probit"))
```

```
summary(myp)
```

- z-test för att testa om varje koefficient för sig är skild från noll, där koefficienternas skattade standardavvikelse ges från den skattade kovariansmatrisen $S[\hat{\beta}]$.
- Wald test kan användas för generella restriktioner på formen $H_0 : \mathbf{R}\beta = \mathbf{q}$ med Walds teststatistika (appr. χ^2 -fördelad med antalet restriktioner som f.g. under nollhypotesen.)

$$W = (\mathbf{R}\hat{\beta} - \mathbf{q})' \{ \mathbf{R} (S[\hat{\beta}]) \mathbf{R}' \}^{-1} (\mathbf{R}\hat{\beta} - \mathbf{q}).$$

- Följande definitioner gäller för Walds teststatistika:

$$\mathbf{R} = \begin{bmatrix} r_{10} & r_{11} & \cdots & r_{1k} \\ r_{20} & r_{21} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n0} & r_{n1} & \cdots & r_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_n \end{bmatrix},$$

där r_{ij} , q_i är reella tal och n är antalet restriktioner under nollhypotesen.

- Detta ger att

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q} \iff \begin{array}{ccccccc} r_{10} \cdot \beta_0 & + & r_{11} \cdot \beta_1 & + & \cdots & + & r_{1k} \cdot \beta_k & = & q_1 \\ r_{20} \cdot \beta_0 & + & r_{21} \cdot \beta_1 & + & \cdots & + & r_{2k} \cdot \beta_k & = & q_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ r_{n0} \cdot \beta_0 & + & r_{n1} \cdot \beta_1 & + & \cdots & + & r_{nk} \cdot \beta_k & = & q_n \end{array}$$

- Skattade koefficienter och kovariansmatris för logit modellen i *R*:
`coef(myl)`, `vcov(myl)`
- Wald test för att både GPA och TUCE inte bidrar som förklaringsvariabler i logit modellen:

`wald.test(b = coef(myl), Sigma = vcov(myl), Terms = 2:3)`

- Likelihood kvottest för allmänna restriktioner under nollhypotesen:

$$LR = -2 [\ln \hat{L}_r - \ln \hat{L}] ,$$

där $\hat{L}_r$ och $\hat{L}$ är log-likelihood funktionerna utvärderade i de respektive restriktade (under nollhypotesen) och icke-restriktade skattade koefficienterna. Sannolikhetsfördelningen för teststatistikan är appr. χ^2 —fördelad med antalet restriktioner som f.g. under nollhypotesen.

- Likelihood kvottest för att både GPA och TUCE inte bidrar som förklaringsvariabler i logit modellen:

```
modelUnRestricted <-  
glm(GRADE~GPA+TUCE+PSI,family="binomial")  
modelRestricted <- glm(GRADE~PSI,family="binomial")  
lrtest(modelUnRestricted,modelRestricted)
```

- Ger testen samma resultat? Varför eller varför inte?
- Dokumentation om detta finns här...

<https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-and-or-similar/>

- Likelihood kvotindexet är ett mått mellan 0 och 1 på hur bra modellen anpassar data:

$$LRI = 1 - \frac{\ln L}{\ln L_0},$$

där $\ln L_0 = n [P \ln P + (1 - P) \ln(1 - P)]$ är log-likelihooden då alla lutningskoefficienter är lika med noll och P är andelen 1:or eller 0:or för responsvariabeln.

- Perfekt anpassning ger $LRI = 1$ och $LRI = 0$ om den skattade modellen inte ens är bättre än en tom modell utan parametrar.
- Ej tolkningsbart mellan 0 och 1.

- Ett beslut tas mellan fler än två alternativ. Vi skiljer mellan ordnade och icke-ordnade beslutsalternativ.
- Exempel på icke-ordnade alternativ: transportsätt till jobbet, val av tvättmedel, val av kandidatprogram, o.s.v.
- Exempel på ordnade alternativ: inställningsfrågor i surveys, grad av anställning, nivå av försäkringsskydd

- Beräkning av multipla integraler för normalfördelningen innebär att en multinomial probit modell inte är särskilt lämplig.
- Multinomial logit modellen kan definieras som

$$P(Y = j) = \frac{e^{\beta_j \mathbf{x}'}}{1 + \sum_{k=1}^J e^{\beta_k \mathbf{x}'}} , \quad j = 1, 2, \dots, J,$$

där β_j är radvektorn av parametrar för kategori j av totalt $J + 1$ kategorier och $\mathbf{x}$ är radvektorn av förklaringsvariabler.

$$P(Y = 0) = \frac{1}{1 + \sum_{k=1}^J e^{\beta_k \mathbf{x}'}} .$$

- Log-likelihooden för multinomiala logit modellen generaliserar den binomiala logit modellen:

$$\ln L = \sum_i \sum_{j=0}^J d_{ij} \ln P(Y_i = j),$$

där $d_{ij} = 1$ om alternativ j väljs av enhet i , och 0 annars.

- Maximum likelihood skattningen fås genom att maximera log-likelihooden numeriskt.
- Log-likelihooden då alla lutningskoefficienter är lika med noll:

$$\ln L_0 = \sum_{j=0}^J n_j \ln P_j.$$

- Koefficienterna är svårtolkade! Marginella effekten från förklaringsvariabel x_k på sannolikheten för kategori j :

$$\frac{\partial P(Y = j)}{\partial x_k} = P(Y = j) \left[\beta_{jk} - \sum_{r=1}^J P(Y = r) \beta_{rk} \right],$$

där $\beta_{0k} = 0$.

- Modellen ger log-odds kvoten för kategorier j och k som

$$\ln \left[\frac{P(Y = j)}{P(Y = k)} \right] = (\beta_j - \beta_k) \mathbf{x}'$$

Exempel: studenter val av studieprogram

- Studenter väljer ett av följande tre studieprogram: akademiskt (academic), allmänt (general) eller yrkesprogram (vocation).
- Deras val kan bero på skrivningspoäng på ett test (write) och deras socioekonomiska status (ses).
- I exemplet på

<https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/>

har data samlats in för variablerna ovan.

- En multinomial logit modell ska anpassas till datamaterialet, se dokumentationen i exemplet.

- Multinomial logit och probit modellerna kan inte hantera ordnade data för den beroende variabeln.
- Linjär regression fungerar inte heller. **Exempel:** survey om inställning till en fråga där svaren rankas och kodas med 0, 1, 2, 3, 4. Linjär regression behandlar skillnaden mellan 4 och 3 p.s.s. som skillnaden mellan 3 och 2, även om värdena bara är en ranking mellan dem.
- Ordnad logit bygger på latent regression. Definiera

$$y^* = \beta \mathbf{x}' + \epsilon,$$

där y^* är icke-observerat, men det gäller att

$$\begin{aligned} y &= 0 \text{ om } y^* \leq \mu_1, \\ &= 1 \text{ om } \mu_1 < y^* \leq \mu_2, \\ &= 2 \text{ om } \mu_2 < y^* \leq \mu_3, \\ &\vdots \\ &= J \text{ om } \mu_J \leq y^*. \end{aligned}$$

- Antag den logistiska fördelningsfunktionen för ϵ , d.v.s.

$P(\epsilon \leq \beta \mathbf{x}') = \Lambda(\beta \mathbf{x}')$. Då följer det att

$$P(y = 0) = P(y^* \leq \mu_1) = P(\beta \mathbf{x}' + \epsilon \leq \mu_1) = \Lambda(\mu_1 - \beta \mathbf{x}'),$$

$$P(y = 1) = \Lambda(\mu_2 - \beta \mathbf{x}') - \Lambda(\mu_1 - \beta \mathbf{x}'),$$

$$P(y = 2) = \Lambda(\mu_3 - \beta \mathbf{x}') - \Lambda(\mu_2 - \beta \mathbf{x}'),$$

$$\vdots$$

$$P(y = J) = 1 - \Lambda(\mu_J - \beta \mathbf{x}'),$$

där $\mu_1 < \mu_2 < \mu_3 < \dots < \mu_J$.

- Log-likelihood funktionen är en generalisering av logit modellen och maximering av log-likelihood funktionen ger skattade värden på $\beta, \mu_1, \mu_2, \mu_3, \dots, \mu_J$.

- Marginell effekt från respektive förklaringsvariabel x_j på respektive sannolikhet för tre kategorier, givet att $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k)$ hålls konstanta:

$$\frac{\partial P(y=0)}{\partial x_j} = -\Lambda(\mu_1 - \beta \mathbf{x}') (1 - \Lambda(\mu_1 - \beta \mathbf{x}')) \beta_j,$$

$$\frac{\partial P(y=1)}{\partial x_j} =$$

$$[\Lambda(\mu_1 - \beta \mathbf{x}') (1 - \Lambda(\mu_1 - \beta \mathbf{x}')) - \Lambda(\mu_2 - \beta \mathbf{x}') (1 - \Lambda(\mu_2 - \beta \mathbf{x}'))] \beta_j$$

$$\frac{\partial P(y=2)}{\partial x_j} = \Lambda(\mu_J - \beta \mathbf{x}') (1 - \Lambda(\mu_J - \beta \mathbf{x}')) \beta_j.$$

- Antag att $\beta_j > 0$. Då gäller att $\frac{\partial P(y=0)}{\partial x_j} < 0$ och $\frac{\partial P(y=2)}{\partial x_j} > 0$.
Marginella effekten på $P(y=1)$ är dock tvetydig!

Exempel: studenter ansökan till *Graduate school*

- Studenter anger hur troligt det är att dom ansöker till *Graduate school*. Responsvariabeln *Apply* antar något av följande alternativ: “*unlikely*”, “*somewhat likely*”, “*very likely*”.
- Deras tilltro kan bero på följande variabler:
 - om minst en av föräldrarna har en universitetsutbildning ($pared = 1$)
 - grundutbildning är på en offentlig *Graduate school* ($public = 1$)
 - genomsnittsbetyg för en student (gpa)
- I exemplet på <https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/> har data samlats in för variablerna ovan.
- En ordnad logit modell ska anpassas till datamaterialet, se dokumentationen i exemplet.